Feature-Detektion, Annotation und Alignment von Metabolomik LC/MS Daten

Dissertation

zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften (Dr. rer. nat.)

der Naturwissenschaftlichen Fakultät III
(Institut für Informatik)
der Martin-Luther-Universität Halle-Wittenberg

vorgelegt von

Ralf Tautenhahn
geb. am 04. März 1976 in Schlema

Halle (Saale), Dezember 2008

Gutachter:
1. Prof. Dr. Stefan Posch
2. Prof. Dr. Sebastian Böcker

Vorgelegt am: 11. Dezember 2008
Datum der Verteidigung: 6. April 2009

Inhaltsverzeichnis

1 Einführung

Der Begriff *Metabolomik*, geprägt in Analogie zu den Begriff Genomik und Proteomik, bezeichnet die Erforschung des Metaboloms eines biologischen Systems. Das *Metabolom* bezeichnet die Gesamtheit der niedermolekularen Substanzen (Metaboliten), die an Reaktionen des Stoffwechsels beteiligt sind. Unterschieden wird hier zwischen dem Primärmetabolismus, d.h. Stoffwechselvorgänge, die für das Wachstum und Überleben eines Systems notwendig sind und dem Sekundärmetabolismus, der alle sonstigen Stoffwechselvorgänge umfasst, welche für das Überleben nicht unmittelbar notwendig sind, aber dennoch im Allgemeinen wichtige Funktionen erfüllen.

1.1 Pflanzenmetabolomik

Beispiele für Stoffwechselvorgänge des Sekundärmetabolismus bei Pflanzen sind Reaktionswege, die Substanzen zum Schutz der Pflanze produzieren, wie z.B. Alkaloide als Fraßschutz oder Carotinoide zum Schutz vor Photooxidation. Substanzen, die zum Primärmetabolismus gezählt werden, sind u.a. Aminosäuren, Zucker, Nucleotide oder Nucleinsäuren. Beispiele für pflanzliche Sekundärmetabolite sind Alkaloide wie Koffein oder Nikotin, Flavonoide wie Kaempferol, Rutin, Biochanin A oder Carotinoide wie β-Carotin oder Lycopin.
Die Gesamtanzahl der im Pflanzenreich vorkommenden Metabolite wird auf bis zu 200.000 geschätzt [Fiehn 2001], die Anzahl der Metabolite in einer Pflanzenart auf bis zu 5000 [Tolstikov et al. 2003]. Da Metabolite chemisch sehr unterschiedlich sind und in einem großen Konzentrationsbereich auftreten, existiert bis zu diesem Zeitpunkt keine analytische Technik, die in der Lage wäre, alle vorkommenden Metaboliten zu detektieren [De Vos et al. 2007]. Innerhalb des letzten Jahrzehnts wurden verschiedene Methoden zur Analyse von Metaboliten in Pflanzenextrakten beschrieben: Gaschromatographie, gekoppelt mit Massenspektrometrie (GC/MS), z.B. [Fiehn et al. 2000, Roessner et al. 2001], Flüssigchromatographie, gekoppelt mit Massenspektrometrie (LC/MS), z.B. [Tolstikov und Fiehn 2002, Roepenack-Lahaye et al. 2004], Kapillar-Elektrophorese, gekoppelt mit Massenspektrometrie (CE/MS), z.B. [Sato et al. 2004], Massenspektrometrie mit Direktinjektion, z.B. [Aharoni et al. 2002, Beckmann et al. 2008] und Kernspinresonanzspektroskopie (NMR), z.B. [Ward et al. 2003]. Einen ausführlichen Überblick über aktuelle Techniken und Anwendungen geben [Dunn 2008] und [Werner et al. 2008].
Die in dieser Arbeit betrachtete LC/MS Technik (Kapitel 2) in Verbindung mit einer Elektrospray-Ionisierung (Abschnitt 2.2) ist in der Lage, einen großen Bereich von polaren bis semipolaren Substanzen zu detektieren. Dazu gehören vor allem Sekundärmetabolite wie Alkaloide, Saponine, Phenolsäuren, Flavonoide, Glucosinolate, Polyamine und Derivate davon, sowie auch einzelne Primärmetabolite wie Aminosäuren.

1.2 Verarbeitung von Metabolomik LC/MS Daten

Die Verarbeitung von Metabolomik LC/MS Daten lässt sich in die folgenden Schritte unterteilen:

1. Zweidimensionale Feature-Detektion und Integration
2. Alignment der korrespondierenden Features über mehrere Messungen
3. Annotation der Features
4. Statistische Analyse, chemische und biologische Interpretation

Ein typisches Metabolomik-Experiment liefert eine zwei- bis dreistellige Anzahl von LC/MS Messungen. Jede dieser Messungen enthält mehrere tausend Features, deren exakte Position und Intensität bestimmt werden muss. Um die Vergleichbarkeit der Intensitäten dieser Features über alle Messungen hinweg zu ermöglichen, ist das Alignment entscheidend und liefert die Basis für die statistische Analyse der Daten. Die Annotation von Features ist ein zusätzlicher Schritt, der sowohl vor als auch nach dem Alignment ausgeführt werden.

1.3 Inhalt dieser Arbeit

Diese Arbeit beschäftigt sich mit den Schritten 1–3 der LC/MS-Datenverarbeitung. In Kapitel 3 wird ein neu entwickelter Algorithmus „centWave“ zur Feature-Detektion von LC/MS-Daten beschrieben. Der Algorithmus kombiniert einen dichteorientierte Methode zur Erkennung von potentiell interessanten Massensignalen mit der kontinuierliche Wavelet Transformation zur Detektion von chromatographischen Peaks. Die Leistungsfähigkeit des Algorithmus wird durch eine Evaluierung und Vergleich mit zwei anderen Feature-Detektions Algorithmen gezeigt.

Kapitel 4 beschreibt eine ebenfalls neu entwickelte Methode zur Gruppierung der von einer Substanz hervorgerufenen Features, verbunden mit der Annotation von Isotopomeren, Addukten und Fragmenten, worüber die Molekülmasse der Substanz bestimmt werden kann. Eine LC/MS-Messung von bekannten Substanzen wurde benutzt, um die Leistungsfähigkeit dieser Methode zu prüfen.

Kapitel 5 beschreibt die Evaluierung von Algorithmen für das Alignment von Metabolomik LC/MS-Daten. Es wurde ein Verfahren entwickelt, um verschiedene Alignment-Algorithmen zu vergleichen und deren Güte zu beurteilen. Mit den erstellten Referenzdatensätzen wurden Vergleiche zwischen vier verschiedenen Alignment-Algorithmen durchgeführt und bezüglich Recall und Precision quantifiziert.

2 Grundlagen der LC/MS

Der Begriff LC/MS steht für die Kopplung von Flüssigchromatographie (liquid chromatography, LC) mit einem Massenspektrometer (MS). Mithilfe der Flüssigchromatographie wird dabei eine Separation des zu analysierenden Substanzgemisches bezüglich bestimmter physikalischer Eigenschaften der Substanzen wie etwa der Polarität durchgeführt. Durch die Kopplung mit einem Massenspektrometer können die eluierenden Substanzen bezüglich ihrer Molekülmasse charakterisiert werden.
Die in dieser Arbeit betrachteten Daten wurden mit der Gerätekombination HPLC/ESI-QTOF-MS gemessen, d.h. Hochleistungsflüssigchromatographie (HPLC), die über Elektrospray Ionisierung (ESI) an ein Quadrupol-Time-of-Flight Massenspektrometer (QTOF-MS) gekoppelt wurde.

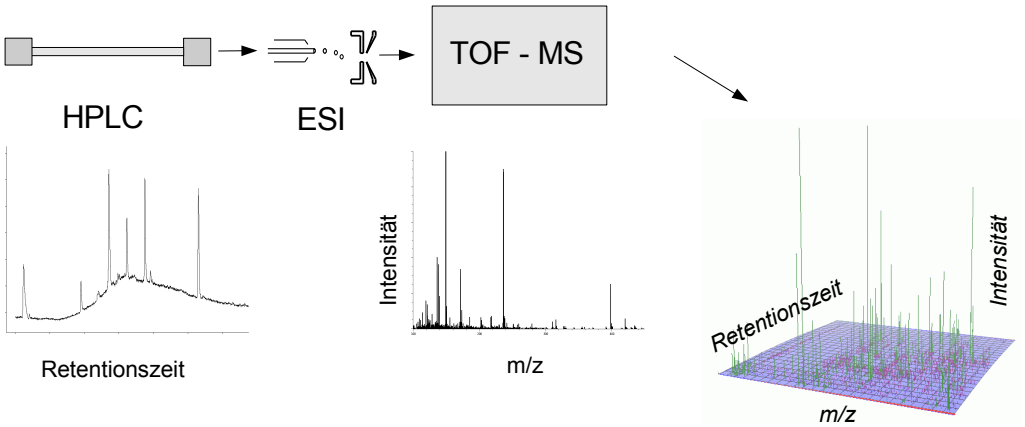

Abbildung 2.1: Struktur eines LC/MS Datensatzes

Abbildung 2.1 illustriert die Struktur eines LC/MS Datensatzes. Die Probe durchläuft die HPLC Säule, wobei verschiedene Verbindungen zu unterschiedlichen Zeiten, der jeweiligen *Retentionszeit*, eluieren. Das Massenspektrometer nimmt mit einer bestimmten Frequenz Massenspektren von dem aus der Säule austretenden Eluat auf. Der resultierende LC/MS Datensatz – der aus vielen aneinandergereihten Massenspektren besteht – hat somit die Dimensionen m/z und Retentionszeit, über denen die Intensität der Signale aus dem Massenspektrum aufgetragen ist.
Die verwendeten Techniken werden im Folgenden näher erläutert.

2.1 Hochleistungsflüssigchromatographie

Allgemein werden als *Chromatographie* physikalische Trennungsmethoden bezeichnet, bei der die zu trennenden Komponenten zwischen zwei Phasen - einer stationären und einer mobilen Phase - unterschiedlich verteilt und dadurch getrennt werden können. Bei der

hier betrachteten *Hochleistungsflüssigchromatographie* ist die mobile Phase eine Flüssigkeit, die unter hohem Druck steht, während die stationäre Phase in eine Säule (ein dünnes Metallrohr) eingebracht wurde. In den in dieser Arbeit beschriebenen Experimenten wur-

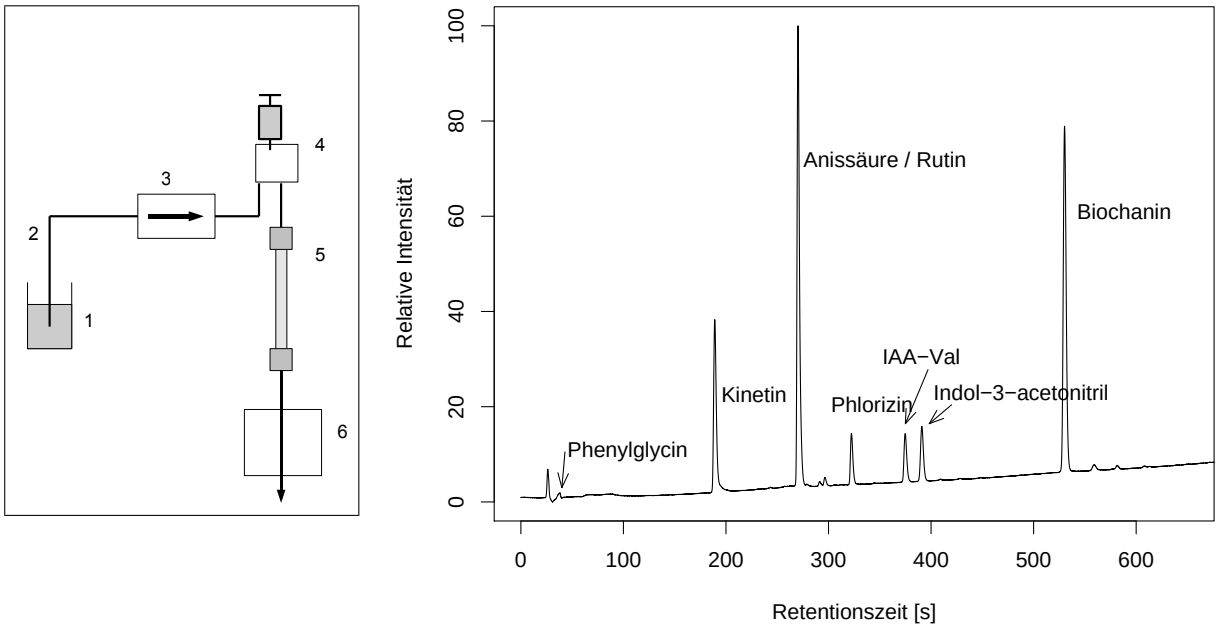

Abbildung 2.2: Links: Schema einer HPLC-Apparatur: 1) Lösungsmittelvorrat, 2) Zuleitung, 3) Pumpe, 4) Probenaufgabe, 5) Säule, 6) Detektor (nach [Meyer 2004]). Rechts: Beispiel für chromatographische Trennung einer Mischung aus acht Verbindungen. Die Detektion wurde mittels UV-Detektor (Absorption bei 254 nm) durchgeführt. Anissäure und Rutin elurieren zur selben Zeit und sind daher als ein gemeinsamer chromatographischer Peak sichtbar.

de die sogenannte Umkehrphasen-Chromatographie benutzt, bei der - im Gegensatz zur klassischen Säulenchromatographie - die stationäre Phase sehr apolar und die mobile Phase relativ polar ist. Je apolarer eine Verbindung, desto höher ist ihre Affinität zur stationären Phase. Apolare Stoffe eluieren daher später als polare. Als stationäre Phase werden bei der als RP C_{18} bezeichneten Technik Kieselgelpartikel benutzt, welche mit Octadecylsilan, einem C_{18} Alkan, funktionalisiert wurden. Als mobile Phase dient ein Acetonitril-Wasser-Gemisch, wobei das Acetonitril/Wasser Verhältnis gesteuert wird („als Gradient gefahren“).
Der linke Teil der Abbildung 2.2 zeigt die wichtigsten Elemente einer HPLC-Apparatur. Die mobile Phase wird unter kontrolliertem Druck zur Säule gepumpt. Die Probe wird aufgegeben und durchläuft die Säule. Am Detektor werden die einzelnen Fraktionen registriert. Eine hierfür gebräuchliche Methode ist die Messung der Lichtabsorption durch die Probe im ultravioletten Wellenlängenbereich (UV-Detektor). Die Darstellung dieser Messung über die Retentionszeit wird als *UV-Chromatogramm* bezeichnet. Abbildung 2.2(rechts) zeigt das UV-Chromatogramm einer Mischung aus acht Verbindungen.

Die eluierenden Verbindungen können direkt zu einem weiterem Gerät, wie z.B. einem Massenspektrometer, geleitet werden. Weiterführende Informationen zur HPLC-Technik finden sich z.B. in [Ardrey 2003, Meyer 2004, Kromidas 2006].

2.2 Elektrospray Ionisierung

Da vom Massenspektrometer nur elektrisch geladene Moleküle gemessen werden können, müssen die zu messenden Analyten der Probe in einen ionisierten Zustand überführt werden. Die Ionisierung des aus der flüssigen Phase eluierenden Analyten erfolgt unter atmosphärischen Druckbedingungen, die dafür verwendeten Techniken werden unter dem Begriff *atmospheric pressure ionisation* (API) zusammengefasst.
Eine gebräuchliche Methode ist neben der *chemischen Ionisierung* unter Atmosphärendruck (atmospheric pressure *chemical ionisation*, APCI) die im Weiteren betrachtete *Elektrospray Ionisierung* (ESI).
ESI ist eine weiche[1] Ionisierungstechnik mit der Ionen aus der flüssigen Phase in die Gasphase überführt werden. Das Grundprinzip ist die pneumatische Zerstäubung einer

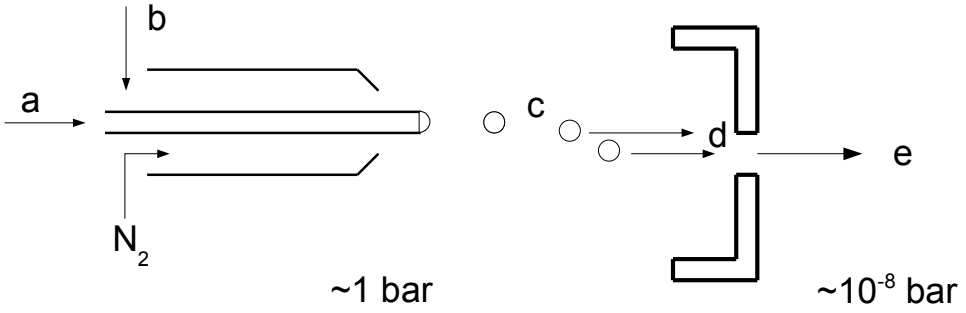

Abbildung 2.3: Schematische Darstellung eines Elektrospray - Interface: (a) Kapillare des Flüssigchromatographen, (b) unter Hochspannung stehende Zerstäuberkapillare, (c) geladene Tröpfchen, (d) Ionen, (e) zum Massenspektrometer (nach [Budzikiewicz und Schäfer 2005]).

Lösung in einem elektrischen Feld. Am Ende der Sprühkapillare bildet sich ein Flüssigkeitskonus, von dem aus ein Nebel von mikrometerkleinen Tröpfchen ins Vakuum sprüht. Durch Verdunstung kommt es zum Schrumpfen der Tröpfchen und mit Abnahme des Tropfenradius sammeln sich Ionen infolge elektrostatischer Abstoßung an der Oberfläche, um schließlich in die Gasphase überzutreten (Abbildung 2.3).
Abhängig von der Polarität, der Konzentration der Substanz, dem verwendeten Lösungsmittel und der Gegenwart von Salzen werden aus einem Molekül M verschiedenste Ionen

[1] Im Gegensatz zur „harten“ Ionisierung werden bei der „weichen“ Ionisierung die Moleküle nicht oder nur geringfügig fragmentiert.

gebildet. Beispiele für im positiven Modus gebildete Ionen sind $[M+H]^+$, $[M+Na]^+$ und $[M+K]^+$. Zusätzlich können Cluster-Ionen wie beispielsweise $[2M+H]^+$, $[2M+Na]^+$ und $[2M+K]^+$ und mehrfach geladene Ionen wie z.B. $[M+2H]^{2+}$ oder $[M+H+K]^{2+}$ auftreten. Weiterhin wird auch die Bildung von Fragmentionen beobachtet, z.B. $[M\text{-}C_6H_9O_5]^+$ bei Glycosiden. Die Zuordnung der gebildeten Ionen, verbunden mit der Rückrechnung auf die Molekülmasse ist Gegenstand von Kapitel 4.

2.3 Massenspektrometrie

Das Grundprinzip der Massenspektrometrie besteht aus

- der Erzeugung von Ionen aus organischen oder anorganischen Verbindungen mittels einer geeigneten Technik,
- der Trennung dieser Ionen aufgrund ihres Masse-Ladungs-Verhältnisses (m/z) sowie
- der qualitativen (m/z) und quantitativen Detektion (relative Menge bzw. Intensität) der jeweiligen Ionen.

Die entsprechenden Teile des Gerätes werden als Ionenquelle, (Massen-)Analysator und Detektor bezeichnet. Die Elektrospray Ionisierung als eine Technik zur Erzeugung von Ionen wurde bereits im vorherigen Abschnitt beschrieben. Für die Trennung und den Nachweis der Ionen stehen verschiedene Analysatoren mit den entsprechenden Detektoren zu Verfügung. Als Beispiele seien das Quadrupol-Massenspektrometer [Miller und Denton 1986], das Ionen-Cyclotron-Resonanz-Spektrometer (ICR, aufgrund der Berechnung des Massenspektrums durch Fourier-Transformation auch als FT-ICR bezeichnet) [Amster 1996] sowie das Time-of-Flight-Massenspektrometer (TOF) [Guilhaus 1995] genannt, dessen Grundprinzip im folgenden Abschnitt dargestellt wird.

2.3.1 Time-of-Flight Massenspektrometrie

Bei der *time-of-flight* (TOF) Massenspektrometrie werden die vorher durch ein elektrisches Feld beschleunigten Ionen aufgrund ihrer unterschiedlichen Flugzeiten entlang einer feldfreien Strecke bekannter Länge getrennt. Leichtere Ionen erreichen den Detektor früher als schwerere Ionen. Die elektrische Ladung q eines Ions der Masse m_i entspricht einer Anzahl z von Elementarladungen e. Die Energieaufnahme E_{el} des Ions bei einer Spannung U ist gegeben durch

$$E_{el} = qU = ezU \tag{2.1}$$

Diese Energie eines geladenen Teilchens in einem elektrischen Feld wird in kinetische Energie umgesetzt :

$$E_{el} = ezU = \frac{1}{2}mv^2 = E_{kin} \tag{2.2}$$

Die Zeit t, die ein Teilchen benötigt, um bei konstanter Geschwindigkeit die Strecke s im feldfreien Flugrohr zurückzulegen, ist gegeben durch

$$t = \frac{s}{v} \tag{2.3}$$

Wird damit v in Gleichung 2.2 substituiert, ergibt sich die Flugzeit

$$t = \frac{s}{\sqrt{\frac{2ezU}{m}}} \tag{2.4}$$

bzw. das Masse/Ladungs-Verhältnis des Teilchens

$$\frac{m}{z} = \frac{2eUt^2}{s^2} \tag{2.5}$$

Die am Detektor registrierten Flugzeiten liefern nach der Umrechnung in m/z Werte das *Massenspektrum*. Die durch die Umrechnung der Flugzeit entstandenen m/z-Werte im Spektrum sind beim TOF-Massenspektrometer (im Gegensatz z.B. zum Quadrupol-Massenspektrometer) nicht äquidistant. Die Menge der in einem (Flugzeit-)Meßintervall registrierten Ionen bestimmt die Intensität eines m/z-Wertes.
Für eine gemessene Ionenspezies entstehen (bedingt durch Flugzeitabweichungen und die bei der Messung auftretenden Toleranzen) mehrere m/z-Werte, die im Massenspektrum dargestellt in etwa eine Gausskurve bilden und als *Peak* bezeichnet werden. Abbildung 2.4 links zeigt drei solcher Peaks. Diese Form der Darstellung des Massenspektrums aus seinen „Rohdaten“ wird als *Profilspektrum* bezeichnet.

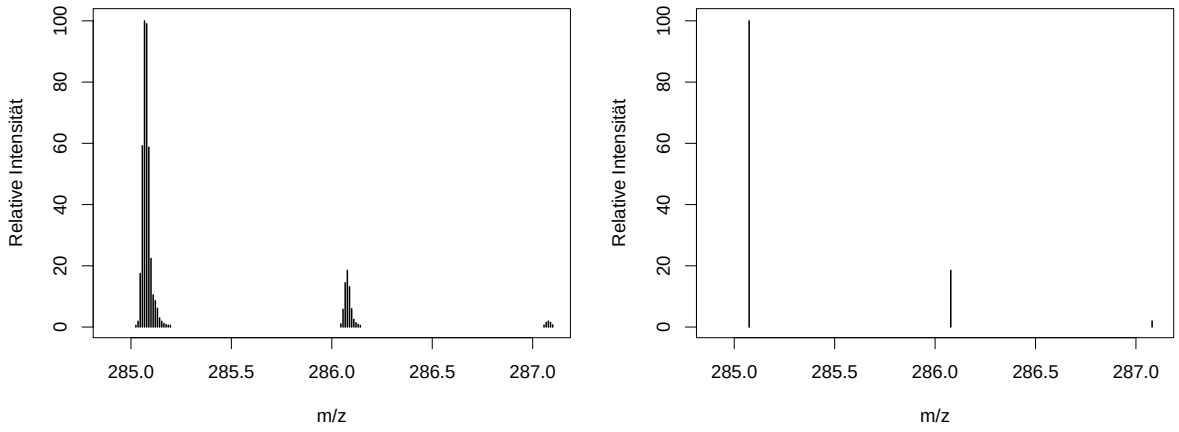

Abbildung 2.4: Profil (links) - und Centroidspektrum (rechts). Dargestellt ist ein Ausschnitt aus dem Massenspektrum von Biochanin A.

Eine weitere, vereinfachte Form der Darstellung des Massenspektrums ist das *Centroid-Spektrum*, bei dem jeder Peak nur noch durch einen m/z-Wert (seinen *Centroid*) repräsentiert wird. Im massenspektrometrischen Sprachgebrauch werden auch die Centroide als

Peak bezeichnet. In dieser Arbeit wird immer dann der Begriff „Centroid“ benutzt, wenn der geschilderte Zusammenhang nur für die Werte aus einem Centroidspektrum gilt, ansonsten wird auch der allgemeinere Begriff „Peak“ verwendet.
Um das Centroid-Spektrum aus dem Profil-Spektrum zu erhalten, werden folgende Schritte durchgeführt:

- die Positionen des Centroide werden aus dem Profilspektrum ermittelt (ein Überblick über entsprechende Methoden sowie ein Beispiel für einen solchen Algorithmus findet sich in [Lange et al. 2006]),
- die Intensitäten der Centroide werden über die maximale Peakhöhe oder aber die Peakfläche ermittelt,
- alle m/z-Werte unterhalb eines bestimmten, vom Benutzer vorgegebenen Schwellwertes, werden als Rauschen betrachtet und verworfen.

Derartig vorverarbeitete Spektren weisen eine Reduktion der Dateigröße um ca. den Faktor zehn auf. Die deutliche Größenreduzierung sowie die leichtere Weiterverarbeitung der Centroid-Spektren begründen die häufige Wahl diese Darstellungsart für die Auswertung und Archivierung. Die Umwandlung wird häufig schon von der Gerätesoftware vorgenommen. Man spricht dabei auch von einer Messung im *Profilmodus* (profile mode) oder *Centroidmodus* (centroid mode).
Weiterführende Information zu Techniken der Massenspektrometrie finden sich z.B. in [Gross 2004, Budzikiewicz und Schäfer 2005, Hoffmann und Stroobant 2001, Cole 1997].

2.3.2 Auflösung und Genauigkeit

Wichtige Kennwerte eines Massenspektrometers sind *Auflösung* (resolution) und *Genauigkeit* (accuracy). Die Massengenauigkeit ist definiert als die Differenz zwischen der gemessenen und der berechneten exakten Masse eines Moleküls und wird normalerweise als relativer Wert in *ppm* (parts per million) angegeben. Moderne TOF-Geräte erreichen eine Massengenauigkeit von ≤ 5 ppm, noch bessere Werte werden von FT-ICR-Geräten erzielt (< 1 ppm).
Der Begriff Auflösung bezieht sich auf die Peakbreite im Profilspektrum und ist definiert als das Verhältnis der gemessenen Molekülmasse[2] m zur Breite des Peaks (Δm) :

$$R = \frac{m}{\Delta m}$$

[2]Bei der Angabe der Auflösung für eine Gerät wird dafür eine Masse genannt, z.B. die eines Kalibrierungs-Peptides.

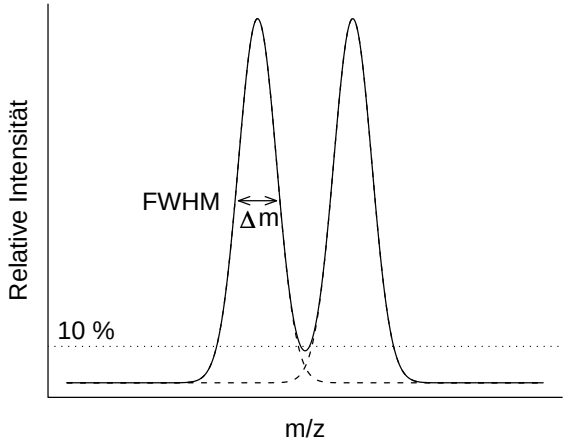

Abbildung 2.5: Verschiedene Definitionen des Auflösungsvermögens: Tal bei 10% der Peakhöhe und Signalbreite bei halber Peakhöhe (FWHM).

Das Auflösungsvermögen eines Gerätes zeigt an, inwieweit Ionen mit nur gering verschiedener Masse getrennt registriert werden können. Ein Auflösungsvermögen von z.B. $R = 5000$ bedeutet, dass beispielsweise einfach geladene Ionen der Masse $m = 5000\ u$ von denen der Masse $m = 4999\ u$ getrennt werden können ($\Delta m = 1$), oder aber Ionen der Masse $m = 100\ u$ von denen der Masse $m = 99.98\ u$ ($\Delta m = 0.02$). Der Begriff „getrennt registriert werden“ ist allerdings nicht einheitlich definiert. Für die Angabe der Auflösung von TOF-Geräten ist es üblich, die Breite des Peaks Δm auf halber Höhe zu betrachten (FWHM), bei anderen Definitionen darf das „Tal“ zwischen zwei gleichhohen, zu trennenden Peaks nicht höher als 10% der Peakhöhe sein (siehe Abbildung 2.5).

2.3.3 Monoisotopische Masse und Isotopomere

Viele natürlich vorkommende Elemente liegen als Gemisch von Isotopen vor. Als *monoisotopische Masse* bezeichnet man die Masse des Moleküls, welches nur aus den am häufigsten vorkommenden Isotopen besteht (z.B. für Biochanin A $[M+H]^+$: $^{12}C_{16}{}^{1}H_{13}{}^{16}O_5^+$, m=285.07575 u). Der entsprechende Peak im Massenspektrum wird als *monoisotopischer Peak* bezeichnet.
Moleküle, die andere Kombinationen von Isotopenspezies enthalten, werden als *Isotopomere* bezeichnet, die gebildeten Peaks als *Isotopenpeaks*. Beispielsweise bezeichnet der ^{13}C-Peak eines Moleküls das Isotopomer, in dem ein ^{12}C-Atom durch ein ^{13}C-Atom ersetzt wurde (z.B. für Biochanin A $[M+H]^+$: $^{13}C_1{}^{12}C_{15}{}^{1}H_{13}{}^{16}O_5^+$, m=286.0791 u). Der Massenunterschied zum monoisotopischen Peak beträgt ≈ 1 u (Genaueres zu den auftretenden Abständen ist in Abschnitt 4.2.2 beschrieben).

Wieviele Isotopenpeaks mit welcher Intensität beobachtet werden ist abhängig von der Konzentration der Substanz und deren chemischer Zusammensetzung. Das theoretische Isotopenmuster lässt sich bei gegebener Summenformel berechnen. Mit zunehmender Größe der Molekülmasse treten immer mehr Isotopenpeaks (immer schwerere Isotopenspezies) auf. In dem bei Metabolomik-Experimenten typischerweise gemessenen Bereich von *m/z* 50 bis 1000 können für die meisten Moleküle die ersten beiden, seltener auch die ersten drei, Isotopenpeaks beobachtet werden. Die Intensitäten der Isotopenpeaks liefern für die spätere Massendekomposition wichtige Hinweise auf die Zusammensetzung des Moleküls [Grange et al. 2006, Böcker et al. 2006, Kind und Fiehn 2007].

2.3.4 Weitere Begriffe

Das Masse-Ladungs-Verhältnis (*m/z*) eines Ions wird angegeben mit m in atomaren Masseneinheiten [u] und der Anzahl der Elementarladungen z (z.B. Biochanin A $[M+H]^+$ *m/z* 285.07575). Bei einfachem Ladungszustand des Ions entspricht der *m/z*-Wert der Masse des Ions, für mehrfache Ladungszustände muss mit z multipliziert werden, um die Masse zu erhalten (siehe Abschnitt 4.2.2).
Die Notation *m/z* wird dem Zahlenwert normalerweise vorangestellt. Gelegentlich findet man die Einheit des Masse-Ladungs-Verhältnisses in *Thomson* (Th) angegeben, wobei dann negativ geladene Ionen mit negativen Werten in Thomson bezeichnet werden (z.B. H^+ : *m/z* 1 = 1 Th; H^-: -1 Th) [Cooks und Rockwood 1991]. Es handelt sich dabei jedoch um keine SI-Einheit.
Für vereinfachte Darstellungen und Notationen wird häufig die *Nominalmasse* des Ions verwendet. Man erhält sie einfach durch das Runden auf ganze Zahlen (z.B. Biochanin A $[M+H]^+$ *m/z* 285).
Abbildung 2.6 zeigt ein Beispiel für ein Massenspektrum.

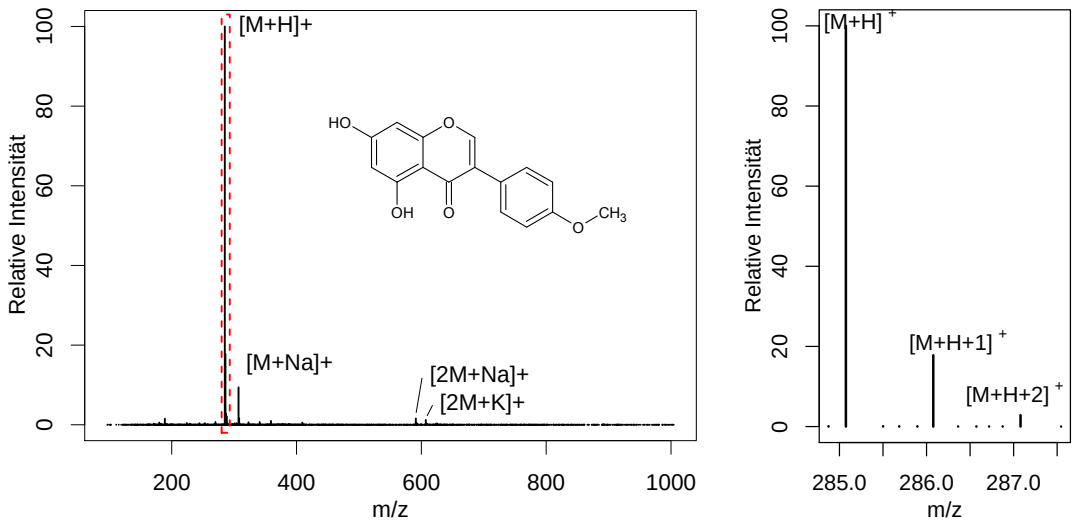

Abbildung 2.6: Links: Massenspektrum von Biochanin A, gemessen nach ESI auf einem Bruker MicrOTOF-Q Massenspektrometer. Die vier Peaks mit der höchsten Intensität sind mit den jeweiligen Ionen gekennzeichnet. Der rot markierte Bereich ist rechts vergrößert dargestellt. Rechts: Monoisotopischer Peak des Ions $[M+H]^+$ von Biochanin A sowie zwei Isotopenpeaks dieses Ions. Die gemessenen m/z Werte dieser Ionen betragen hier 285.0739, 286.0773 und 287.0794.

2.4 LC/MS Daten

Wie in der Einleitung dieses Kapitels erwähnt, besteht der aus einer LC/MS-Messung resultierende Datensatz aus vielen aneinandergereihten Massenspektren (auch als Scans bezeichnet), die mit einer bestimmten Frequenz (der *Scanfrequenz*) von dem aus der Säule austretenden Eluat gemessen wurden. Das von einer Substanz gemessene Massenspektrum erscheint im LC/MS Datensatz daher mehrfach, wie oft ist abhängig von der Scanfrequenz und der Chromatographie. Die Intensität des Spektrums wird von der Konzentration der Substanz zum Aufnahmezeitpunkt bestimmt.
Abhängig von der Substanz und der chromatographischen Trennung zeigt der Konzentrationsverlauf dabei in etwa die Form einer Gausskurve. Das von einem Ion somit hervorgerufene zweidimensionale Signal wird als *Feature* bezeichnet. Abbildung 2.7 zeigt die von Biochanin A $[M+H]^+$ und zwei seiner Isotopomere gebildeten Features in einem LC/MS Datensatz.

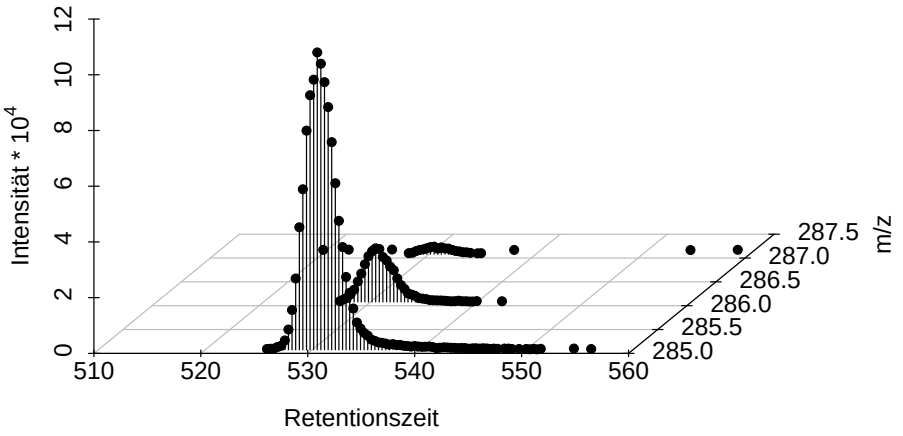

Abbildung 2.7: Ausschnitt aus einem LC/MS Datensatz. Dargestellt sind die durch Biochanin A $[M+H]^+$ und zwei seiner Isotopomere hervorgerufenen Features. Ein Schnitt durch die Retentionszeit-Achse entspricht wieder einem einzelnen Massenspektrum (vgl. Abb. 2.6 rechts).

2.4.1 Massensignal und EIC

Betrachtet man die gemessenen Werte aus einem LC/MS-Datensatze in einem schmalen m/z-Intervall über die gesamte oder einen Teil der Retentionszeit, so wird dies als *Massensignal* bezeichnet. Die zeitliche Darstellung der über m/z summierten Intensitätswerte für diesen Intervall wird als *Extrahiertes Ionen Chromatogramm* (extracted ion chromatogram, EIC) bezeichnet[3].

[3]Gelegentlich findet man dafür auch die Bezeichnungen XIC oder SIC (selected ion chromatogram).

Abbildung 2.8 zeigt Beispiele für die Darstellung eines Massensignals.
Das EIC gibt einen Überblick über die Chromatographie bezogen auf ein m/z-Intervall. Das von einem Ion hervorgerufene Feature stellt sich im EIC als *chromatographischer Peak* dar. Die Darstellung als EIC wird benutzt um

- nach bekannten Massensignalen zu filtern und diese darzustellen,
- chromatographische Peaks zu erkennen und
- diese zu quantifizieren.

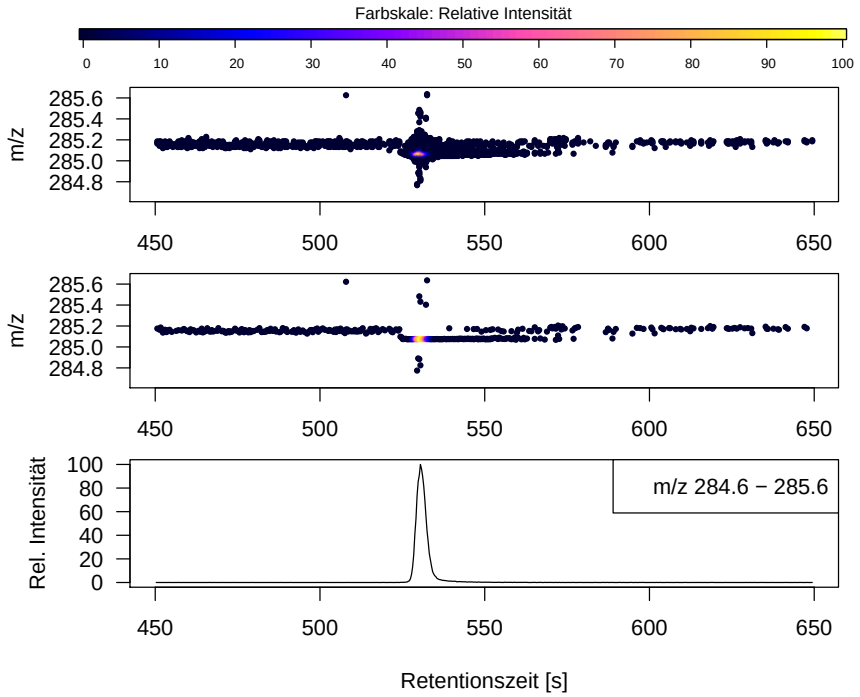

Abbildung 2.8: Massensignal im Profilmodus (oben) und im Centroidmodus (Mitte) von Biochanin A $[M+H]^+$, dargestellt im Bereich von m/z 284.6-285.6 und 450-650 s mit farblich kodierten Intensitätswerten. Unten: Darstellung des Intensitätsverlaufs für diesen Bereich als Extrahiertes Ionen Chromatogramm (EIC).

2.4.2 Total Ionen Chromatogramm

Projiziert man nicht mehr nur einen kleinen Intervall, sondern sämtliche vorkommenden m/z-Werte auf die Zeitachse, so erhält man das *Total Ionen Chromatogramm* (total ion chromatogram, TIC). Das TIC (Abbildung 2.9 oben) gibt einen Überblick über

die Chromatographie bezogen auf alle Massensignale. Es ist dazu geeignet, einen Überblick über den gemessenen Datensatz zu erhalten, vergleichbar in etwa mit einem UV-Chromatogramm (vgl. Abb. 2.2). Einige Arten von Problemen bei der Messung, wie beispielsweise deutliche Retentionszeitsverschiebungen sowie stärkere Verunreinigungen, können dabei erkannt werden.

Abbildung 2.9 stellt ein Beispiel für TIC und EIC dar, gezeigt für zwei Substanzen mit selber Retentionszeit.

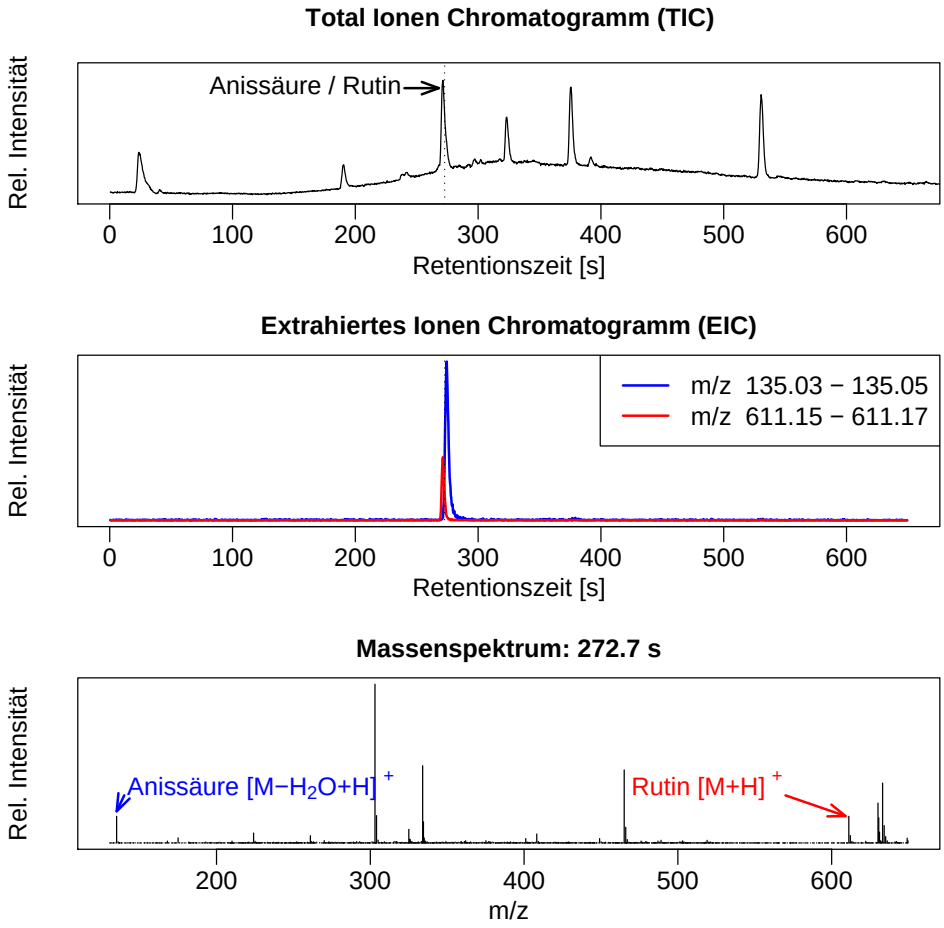

Abbildung 2.9: Beispiel für Total Ionen Chromatogramm und extrahierte Ionen Chromatogramme. Der Strich im obersten Bild markiert die Stelle, für die das entsprechende Massenspektrum (unterstes Bild) dargestellt ist. Für zwei Ionen –aus zwei chromatographisch nicht getrennten Substanzen, Rutin und Anissäure– sind die extrahierten Ionen Chromatogramme (Mitte) dargestellt. Während die Substanzen im TIC nicht getrennt werden, lassen sie sich im Massenspektrum (und damit auch über die EIC) klar unterscheiden.

2.4.3 Zusammenfassung

Durch die Verbindung von Hochleistungsflüssigchromatographie mit einer hochauflösenden Massenspektrometrie entsteht eine sehr leistungsfähige Analysetechnik, die die Erfassung einer Vielzahl von Substanzen innerhalb komplexer Stoffgemische, wie z.B. pflanzlicher Extrakte, möglich macht. Die Komplexität der entstehenden Daten – eine einzige Messung liefert ca. 50 - 500MB – erfordert eine weitgehend automatisierte Verarbeitung, die hohen Ansprüchen in Bezug auf Genauigkeit, Reproduzierbarkeit und nicht zuletzt Geschwindigkeit gerecht werden muss.

3 Feature-Detektion

Die Feature-Detektion stellt einen entscheidenden Schritt im Prozess der LC/MS Datenverarbeitung dar. Von ihrer Zuverlässigkeit und Genauigkeit ist abhängig, wie gut die nachfolgenden Verarbeitungsschritte – Annotation und Alignment – funktionieren. Die exakte Lokalisierung und Quantisierung aller Features entscheidet über die Qualität der abschließenden statistischen Analyse des Experiments.
Die genauen Aufgaben der Feature-Detektion sind

1. die Erkennung aller in einer LC/MS Messung enthaltenen Features,

2. die Bestimmung der Feature-Koordinaten in m/z und der Retentionszeit,

3. die Quantifizierung der Features, sowie

4. die Ableitung von qualitativen Merkmalen (z.B. Signal-Rausch-Verhältnis) für jedes Feature.

Besondere Herausforderungen für den Detektions-Algorithmus bestehen darin, auch Features mit geringer Intensität –hervorgerufen durch Substanzen sehr niedriger Abundanz oder geringer Ionisierbarkeit– zu erkennen und andererseits feature-ähnliche Signale, die durch z.B. sogenanntes chemisches Rauschen verursacht werden, zu vermeiden.
Ein neuer Algorithmus („centWave“) zur Feature-Detektion wurde entwickelt, der im Vergleich mit zwei anderen Algorithmen deutlich bessere Ergebnisse in Bezug auf diese Anforderungen erreicht. Das Verfahren ist weitestgehend universell und wurde erfolgreich auf den Messungen von verschiedensten LC/MS Gerätekombinationen, u.a. HPLC/QTOF, UPLC/QTOF, HPLC/Orbitrap, aber auch CE/MS und GC/MS, getestet. Einzige Voraussetzung ist das Vorliegen der Daten im Centroid-Modus. Die Speicherung in diesem Format erfolgt bei vielen hochauflösenden Massenspektrometern standardmäßig, schon aufgrund des anwenderfreundlicheren Datenvolumens (z.B. $\approx$ 60 MB im Centroid-Modus statt 600 MB im Profilmodus für eine LC/MS Messung beim Bruker MicrOTOF-Q). In anderen Fällen können Programme wie die *OpenMS TOPP Tools* [Kohlbacher et al. 2007] zur Konvertierung eingesetzt werden. Die Centroidisierung mittels der Gerätesoftware bietet jedoch den Vorteil, dass dabei gerätespezifische Modelle verwendet werden, was die bestmögliche Massengenauigkeit liefern sollte.
Der centWave-Algorithmus wurde in [Tautenhahn et al. 2008] beschrieben und ist seit Mitte 2007 im Framework XCMS (`http://www.bioconductor.org/packages/bioc/html/xcms.html`) verfügbar.

3.1 Allgemeines und Begriffe zur Feature Detektion

Durch die Art und Weise der Entstehung einer LC/MS Messung – Massenspektren gemessen über den Zeitverlauf einer chromatographischen Trennung – unterscheiden sich dessen Dimensionen, *m/z* und Retentionszeit, in vielerlei Hinsicht.

- Skala: Während die Meßzeitpunkte in der Retentionszeit bei den hier betrachteten Daten aufgrund der konstanten Scanfrequenz diskret und äquidistant sind, ist die Skala der *m/z*-Werte im Spektrum, entstanden durch die Umrechnung der Ionen-Flugzeiten, kontinuierlich. Ein solcher LC/MS Datensatz lässt sich daher nicht ohne weiteres in eine diskrete Matrix-Repräsentation überführen.

- Ausdehnung eines Features: Die Breite eines Features in der Retentionszeit (RT) beträgt – abhängig von Einstellungen der Chromatographie – typischerweise zwischen 5 und 10 Sekunden. Die Ausdehnung bzw. Schwankung der Centroide, aus denen ein Feature besteht, beträgt bei modernen Massenspektrometern weniger als 20 ppm. Das entspricht bei einem Feature bei *m/z* 500 einer Schwankung von weniger als *m/z* 0.01.

- Abweichungen zwischen LC/MS Datensätzen vom selben Gerät beim Vergleich mehrerer Messungen : RT zwischen 5 und 10 Sekunden, *m/z* weniger als 0.05

Aufgrund der unterschiedlichen Charakteristik der beiden LC/MS Dimensionen werden diese im Allgemeinen bei der Verarbeitung verschieden behandelt. Häufig genutzt werden insbesondere die im Gegensatz zur Retentionszeit deutliche geringeren Abweichungen in *m/z*-Richtung um eine Diskretisierung in *m/z* durchzuführen, was die Verarbeitung vereinfacht und beschleunigt. Für die Feature Detektion wird daher typischerweise ein zweistufiges System verwendet:

- **Separation von Massensignalen**
  Im ersten Verarbeitungsschritt werden nur die *m/z* Werte im Zeitverlauf betrachtet. Ziel hierbei ist es, etwaige Abweichungen in *m/z* auszugleichen und gleichzeitig hinreichend „dünne“ Massensignale zu separieren, die dann im nächsten Schritt auf chromatographische Peaks untersucht werden können.

- **Chromatographische Analyse**
  Betrachtet man nun die Intensitätswerte eines einzelnen Massensignals (in der Regel in Form von EIC's), so können chromatographische Peaks erkannt werden. Die Lokalisation und Integration dieser chromatographischen Peaks (Abb. 3.1 unten) liefert schließlich die einzelnen LC/MS Features.

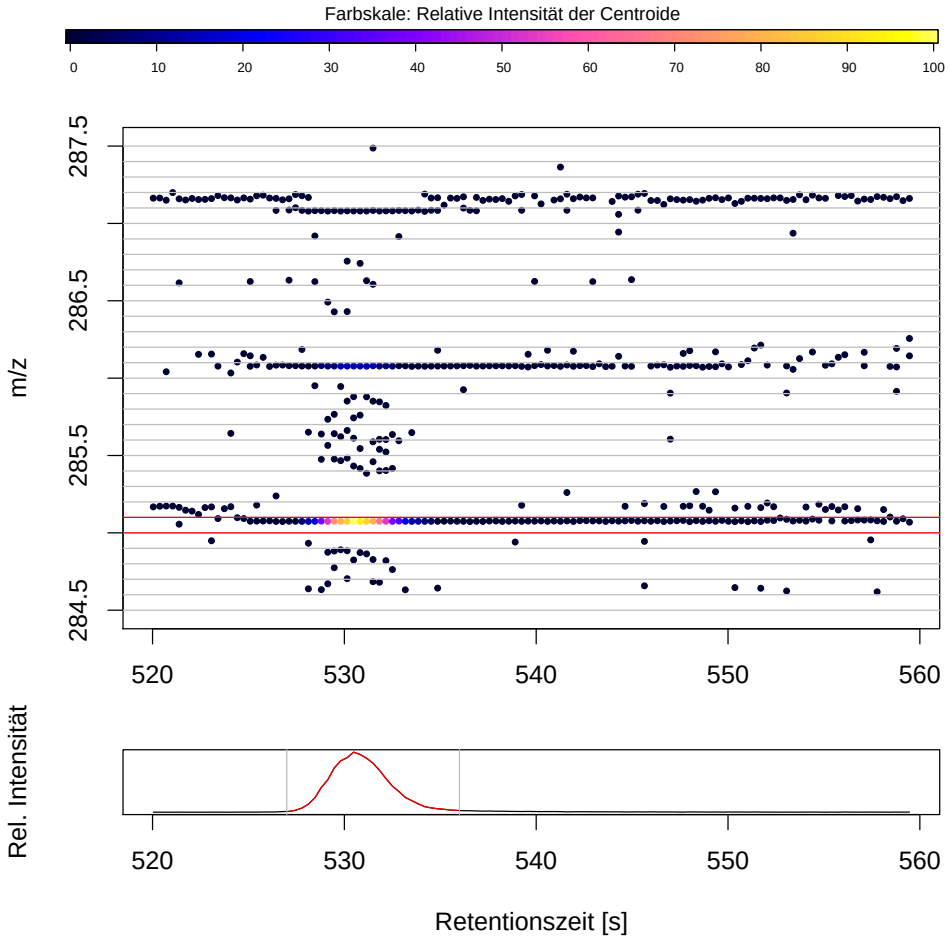

Abbildung 3.1: Typische LC/MS Datenverarbeitung: Diskretisierung von Massensignalen durch *Binning* (Breite 0.1 m/z, oben) und chromatographische Analyse mit Integration der Peakintensität (gezeigt für m/z 285.0 - 285.1, unten).

3.2 Herkömmliche Verfahren

3.2.1 Verfahren zur Separation von Massensignalen

Binning

Ein klassisches und häufig angewendetes Verfahren zur Separation von Massensignalen ist das sog. *Binning*, bei dem die m/z-Achse in äquidistante Intervalle (Standardwert 0.1 m/z) eingeteilt (Abb. 3.1 oben) und die ursprünglichen Werte auf diese Intervalle abgebildet werden. Ergebnis dieser Prozedur ist eine Matrix, die zu den Dimensionen Retentionszeit und m/z die jeweiligen Intensitäten enthält. Eine Zeile dieser Matrix (alle Retentionszeiten für ein m/z-Intervall) wird dann als Massensignal betrachtet.

Der Binning-Ansatz ist vielseitig verwendbar, schnell und leicht zu implementieren. Das Verfahren hat aber auch einige Nachteile, die bereits in [Daszykowski und Walczak

2006, Smith et al. 2006, Stolt et al. 2006, Aberg et al. 2008] erwähnt werden. Das wesentliche Problem ist die Wahl der Intervallbreite (bin size) beim Binning. Auf der einen Seite muss die Intervallbreite groß genug sein, um die gerätebedingten Schwankungen in m/z auszugleichen und alle zu einem jeweiligen chromatographischen Peak gehörenden Datenpunkte zu beinhalten. Ist sie zu klein gewählt, enthält das resultierende Massesignal zu wenig Datenpunkte und die charakteristische Peakform im Chromatogramm geht verloren. Auf der anderen Seite darf die Intervallbreite auch nicht zu groß gewählt werden, da dann im Datensatz eigentlich getrennte Massensignale zusammenfallen. Im Idealfall können darin auftretende Features bei der nachfolgenden chromatographischen Analyse zwar wieder getrennt werden, dies ist jedoch nur dann möglich, wenn diese Features nicht zur selben Zeit auftreten. Desweiteren erschwert das bei breit gewählten Intervallen verstärkt sichtbare chemische Rauschen die Erkennung chromatographischer Peaks, da für die Bildung der EIC's dann wesentlich mehr Intensitätswerte aufsummiert werden.
Ein weiteres Problem ist gegeben durch die beim TOF-Massenspektrometer auftretende konstante Auflösung R, $R = \frac{m}{\Delta m}$. Bei Messungen im Profil-Modus zeigt sich dies dadurch, dass Features im hohen Massenbereich eine größere Ausdehnung in m/z aufweisen, als Features im niedrigen Massenbereich. Im Centroid-Modus weisen die Centroide im höheren Massenbereich entsprechend stärkere Schwankungen auf. Demzufolge sollte eigentlich auch die Intervallbreite bei steigendem m/z-Wert zunehmen, was in den Implementierungen jedoch nicht üblich ist.

„Potential field“ [Stolt et al. 2006]
Eine Alternative zur verbreiteten Binning-Technik, geeignet für centroidisierte Daten, wurde in [Stolt et al. 2006] vorgestellt. Die Autoren legen die Beobachtung zugrunde, dass Bereiche in denen ein deutliches chromatographisches Signal auftritt, Bereiche hoher Dichte sind, die jeweils von einem leeren Bereich(„specific data void“) auf der m/z-Achse umgeben sind. Begründet wird dies damit, dass die entstehende „Lücke“ um ein solches Signal bei der Konvertierung der Spektren vom Profil- in den Centroid-Modus entsteht und außerdem die Wahrscheinlichkeit, zwei oder mehr Centroide mit ähnlichem m/z-Wert bei derselben Retentionszeit zu finden, gering ist. Um solche Regionen zu detektieren, wird für jeden Centroid im Datensatz sowohl die Distanz zum nächsten Centroid im selben Spektrum als auch die m/z-Distanz zu dem nächstgelegen Centroid im darauffolgenden Spektrum bestimmt. Beide Distanzen werden zu einem „potential value“ kombiniert. Die Gesamtheit dieser Wert wird als „potential field“ bezeichnet, aus dem nach Glättung und Schwellwertbestimmung dann in m/z und RT beschränkte Massensignale bestimmt werden. Ergebnis dieser Prozedur ist eine Matrix, die nur die Regionen der chromatographischen Peaks mit ihren spezifischen Massenbereichen enthält und alle anderen Werte auf Null gesetzt sind. Eine weitergehende genaue Analyse bezüglich der Intensitäten der einzelnen Massensignale wird bei diesem Verfahren nicht durchgeführt. Die Laufzeit bei

diesem Verfahren ist erheblich, angegeben sind $\sim$ 2h für eine LC/MS Messung (ohne CPU-Angabe).

Kalman Filter [Aberg **et al. 2008**]
Ein weiterer Ansatz benutzt Kalman-Filter um sogenannte „pure ion chromatograms“ zu extrahieren. Der Kalman-Filter [Kalman 1960] ist ein Algorithmus zur Zustandsschätzung dynamischer Systeme, der in vielen technischen Anwendungen wie z.B. Radarsystemen, Navigationssystemen oder dem Motion Tracking in der Bildverarbeitung verwendet wird. Die Autoren vergleichen den Verlauf eines LC/MS-Massensignals mit der auf einem Radarschirm beobachteten Bewegung von Objekten. Beide Signaltypen werden als strukturierte Signale („structured signals“) bezeichnet.
Da sowohl die aufeinanderfolgenden m/z-Werte eines deutlichen Massensignals aber auch dessen Intensitätswerte bestimmte strukturierte Eigenschaften besitzen (geringe Abweichungen in m/z bzw. ein erst an- und dann wieder absteigendes Signal in der Intensität) werden für jedes Massensignal für einen Tracker zwei parallele Kalman-Filter benutzt, je einer für m/z und Intensität. Der Algorithmus wird in der Richtung vom letzten zum ersten Scan durchgeführt, da intensive chromatographische Peaks ein Tailing aufweisen und deswegen einfacher für den Tracker zu verfolgen seien. Das Ergebnis ist eine Liste von „pure ion chromatograms“, in m/z und RT beschränkte Massensignale mit jeweils einer exakten Schätzung für den m/z-Centroid dieser Region. Auch bei diesem Verfahren verzichten die Autoren auf eine weitergehende genaue Analyse bezüglich der Intensitäten der einzelnen Massensignale, obwohl auch in den verwendeten Abbildungen der Veröffentlichung zu sehen ist, dass etliche der erkannten Regionen mehr als einen chromatographischen Peak aufweisen. Die Laufzeit wird in [Aberg et al. 2008] mit rund 5 Minuten pro LC/MS Messung angegeben (ohne CPU-Angabe).

3.2.2 Chromatographische Analyse mittels angepasster Filter

Im Folgenden werden die Grundlagen der Technik des *angepassten Filter*, sowie dessen Anwendung auf die Analyse chromatographischer Peaks vorgestellt.
Sei $s(t)$ ein zeitlich begrenztes Nutzsignal mit bekannter Charakteristik, welches überlagert wird von einem Rauschsignal $n(t)$. Das beobachtete Signal x ergibt sich als Summe von Nutz- und Rauschsignal : $x(t) = s(t) + n(t)$. Die Anwendung eines Filters mit der Funktion h als Impulsantwort auf das Signal x ergibt das Ausgangssignal:

$$y(t) = x(t) * h(t) = (s * h)(t) + (n * h)(t) = y_s(t) + y_n(t) \ . \tag{3.1}$$

Gesucht wird nun der als *angepasster Filter* (*matched filter*) bezeichnete Filter h, welcher das Verhältnis der Leistung von Nutzsignalanteil y_s gegenüber dem Störsignalanteil y_n

am gefilterten Signal maximiert :

$$SNR = \frac{|y_s|^2}{E\left(|y_n|^2\right)} \ . \tag{3.2}$$

SNR wird als Signal/Rausch-Verhältnis bezeichnet. Die Übertragungsfunktion $H(f)$ des angepassten Filters h wird berechnet als

$$H(f) = \frac{S^*(f)}{P_n(f)} \ , \tag{3.3}$$

wobei $S^*(f)$ die komplex konjugierte des Fourier-transformieren Signals $S(t)$ und $P_n(f)$ die Rauschleistung ist. Die Rauschleistung $P_n(f)$ berechnet sich als die Fourier-transformierte der Autokorrelationsfunktion $R_n(t)$ des Rauschens $n(t)$.
Bei Annahme von weißem (unkorreliertem) Rauschen lässt sich zeigen, dass das Verhältnis in (3.2) maximal wird, wenn die Impulsantwort $h(t)$ des gesuchten angepassten Filter der zeitgespiegelten Nutzsignalfunktion $s(t)$ entspricht [Kiencke et al. 2008].

Angepasste Filter werden in der Chromatographie seit den 1970er Jahren verwendet [van Rijswick 1974]. Die Methode wurde später verfeinert und auch auf LC/MS Daten angewandt [van den Bogaert et al. 1994, Zhang und McElvain 1999, Danielsson et al. 2002, Andreev et al. 2003, Wang et al. 2006, Wang et al. 2008].
Sowohl [Danielsson et al. 2002] als auch [Andreev et al. 2003] überführen die LC/MS Daten mittels Binning in eine Matrixrepräsentation, und wenden einen *angepassten Filter* (*matched filter*) an, um die Daten in chromatographischer Richtung zu filtern, was die nachfolgende Feature Detektion erleichtert. In [Danielsson et al. 2002] werden die Effekte bei der Anwendung einer Gauss-Funktion sowie der zweifach abgeleiteten Gauss-Funktion als angepasster Filter untersucht. Die Autoren zeigen, dass die Gauss-Funktion als angepasster Filter verwendet werden kann, wenn das zu erwartende Signal die Form einer Gauss-Funktion besitzt und nur durch weißes Rauschen überlagert wird. Die zweifach abgeleitete Gauss-Funktion ist hingegen dann besonders gut als Filter geeignet, wenn das zu untersuchende Signal zusätzlich auf einer erhöhten Basislinie liegt, was bei chromatographischen Peaks recht häufig der Fall ist (siehe Abbildung 3.2). Diese erhöhte Basislinie lässt sich als niederfrequente Störung interpretieren und wird bei Verwendung der zweifach abgeleitete Gauss-Funktion – die als Filter eine Bandpass-Charakteristik besitzt – entfernt. Das Signal/Rausch-Verhältnis wird jedoch nicht in dem Maße erhöht, wie bei Verwendung der nicht abgeleiteten Gauss-Funktion. Desweiteren wird gezeigt, dass beide der untersuchten angepassten Filter relativ unempfindlich (im Bezug auf die Verbesserung des Signal/Rausch-Verhältnisses) gegen abweichende Breite sowie Peakform des zu untersuchenden Signalpeaks sind. Die Autoren zeigen an ausgewählten LC/MS Daten, dass die Anwendung der zweifach abgeleiteten Gauss-Funktion als angepasster Filter ein geeigneter Vorverarbeitungsschritt für die Peak-Erkennung in Chromatogrammen ist.

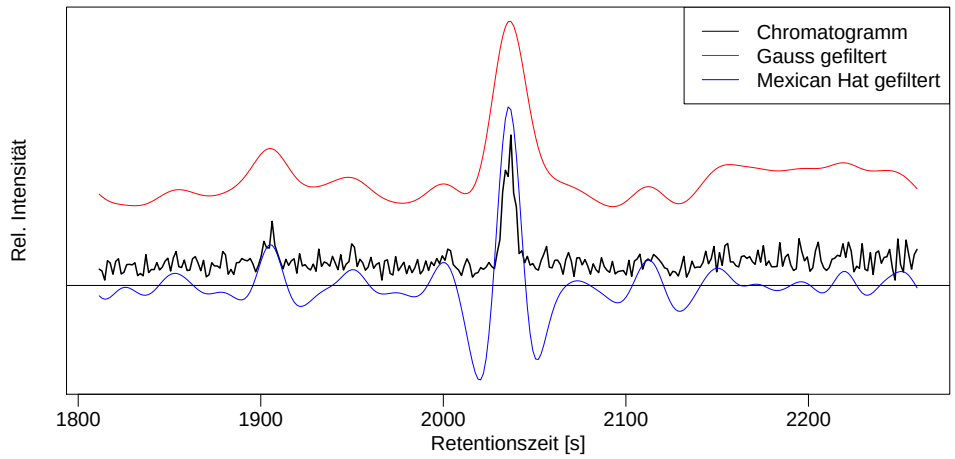

Abbildung 3.2: Chromatogramm aus einem LC/MS Datensatz und Anwendung einer Gauss-Funktion sowie einer zweifach abgeleiteten Gauss-Funktion („Mexican Hat“) als angepasstes Filter.

Die Autoren von [Andreev et al. 2003] bezeichnen ihr Verfahren mit MEND („matched filtration with experimental noise determination“). Verwendet wird ebenfalls eine Gauss-Funktion als Peakmodell. Im Gegensatz zum sonst häufig verwendeten weißen oder farbigem Rauschen als Rauschmodell, wird bei diesem Verfahren die Rauschleistung aus Teilen der verwendeten Daten geschätzt und zur Berechnung des angepassten Filters (Gleichung 3.3) benutzt. Die Rauschleistung wird nach Angabe der Autoren aus „leeren“ Chromatogrammen berechnet – wobei unklar bleibt, wie diese ausgewählt werden. Insbesondere lässt sich die Tatsache, ob ein Chromatogramm „leer“ ist, mit einiger Sicherheit eigentlich erst *nach* erfolgter Peak Detektion feststellen.
Mittels der experimentell bestimmten Übertragungsfunktion wird anschließend die angepasste Filterung für alle Chromatogramme durchgeführt. Auf der nunmehr gefilterten LC/MS Matrix wird anschließend die Peak Detektion durchgeführt, allerdings ausschließlich in m/z Richtung, d.h. die chromatographische Peak-Information ist für die Autoren nicht von Bedeutung. Beschrieben ist ein Vergleich bezüglich der ermittelten m/z-Werte einiger ausgewählter Substanzen sowie Abständen bei Isotopenmustern. Dabei zeigte sich der MEND-Algorithmus vorteilhaft gegenüber einfacher Mittelung von zehn Spektren, Gauss- sowie zweifach abgeleitetem Gauss-Filter.

Probleme bei Verwendung des angepassten Filters

Der angepasste Filter ist eine etablierte Methode in der chromatographischen Analyse und wurde auch weiter adaptiert [Andreev et al. 2003, Wang et al. 2008]. Das Verfahren ist besonders gut geeignet, wenn die interessierenden Signale nur eine eng begrenzte

Frequenzbreite besitzen. Für die Chromatogramme von LC/MS Daten wird dies im Allgemeinen angenommen, vor allem dass die Breite eines Peaks im Chromatogramm nur eine geringe Variabilität besitzt. Die Praxis zeigt jedoch, dass einige Substanzen aufgrund von stärker ausgeprägten Wechselwirkungen mit der stationären Phase gegenüber dem Normalfall deutlich breitere chromatographische Peaks erzeugen oder aber auch Peaks mit einem ausgeprägten Tailing gebildet werden. Abbildung 3.3 zeigt beispielhaft für einen

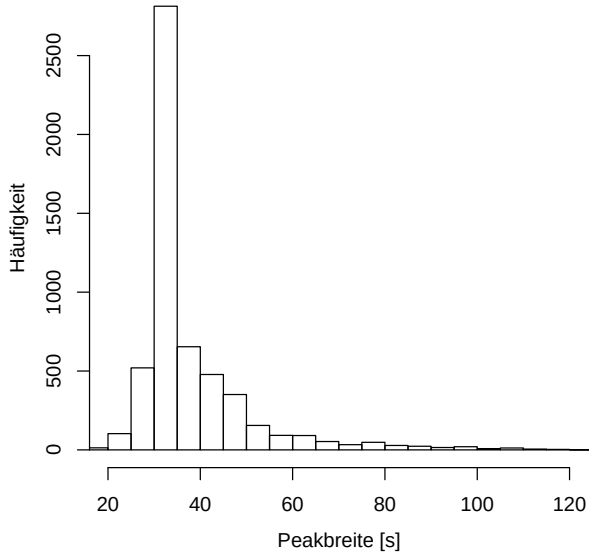

Abbildung 3.3: Verteilung der chromatographischen Breite von 5533 Features aus einer LC/MS Messung eines *A.thaliana* Blattextrakts (Datensatz E105, Details im Anhang). Gezeigt ist die vom centWave Algorithmus erkannte Peakbasisbreite. Zur Feature-Detektion wurden die Parameter snthr=6, ppm=80, peakwidth=(20,50) benutzt.

Datensatz, dass Peaks mit einer Basisbreite von 30-35 Sekunden hier besonders häufig sind, jedoch treten auch Peaks mit einer Breite von 60 Sekunden und mehr auf. Soll zur Verarbeitung ein angepasster Filter für diesen Datensatz verwendet werden, stellt sich somit die Frage nach der optimalen Breite des Modellpeaks. Abbildung 3.4 zeigt die sich ergebenden Probleme bei dieser Parameterwahl. Dargestellt sind zwei ausgewählte Chromatogramme aus derselben Messung, die auch für Abbildung 3.3 verwendet wurde. Im oberen Chromatogramm sollte der Parameter σ des Modellpeaks bei ca. 5 Sekunden gewählt werden, um die dicht aufeinanderfolgenden, schmalen chromatographischen Peaks zu erkennen. Für die Erkennung der drei Peaks im unteren Chromatogramm wäre hingegen ein deutlich breiter gewähltes σ von ca. 20 Sekunden optimal.
Durch numerische Simulation wurde in [Danielsson et al. 2002] gezeigt, dass die optimale Breite des Modellpeaks für den angepassten Filter beim 2 bis 2.5-fachen der Signalbreite liegt. Weiterhin zeigte die Simulation, dass die Signal/Rausch Verbesserung für Signale,

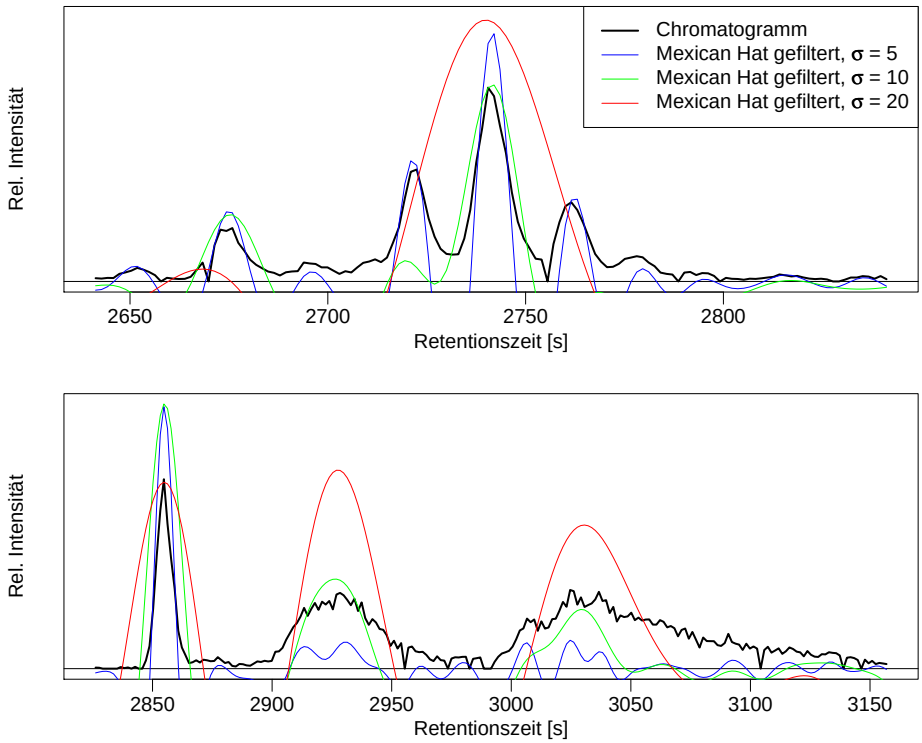

Abbildung 3.4: Zwei verschiedene Chromatogramme aus einer Messung des Datensatzes E105 mit Anwendung des Mexican Hat als angepasster Filter. Die negativen Filteranteile wurden für die Darstellung weggelassen.

die weniger als die 1.5-fache Breite des Modellpeaks besitzen, sehr rasch abnimmt und für Signale die mehr als dreimal so breit sind, langsam abnimmt.

Die Wahl der Breite des Modellpeaks ist bei diesem Verfahren folglich eine kritische Größe und beeinflusst das Ergebnis der Feature Detektion in erheblichem Maße.

3.2.3 Verfügbare Programme zur Feature-Detektion

Neben kommerziellen Programmen zur Feature Detektion von LC/MS Daten wie MetAlign [Tikunov et al. 2005], MarkerLynx (Waters), MarkerView (ABI/Sciex) existieren eine Reihe weiterer Programme, die häufig nur speziell für Proteomik Daten konzipiert wurden. Ein Überblick findet sich in z.B. in [Katajamaa und Oresic 2007].

Es existieren momentan nur zwei für Metabolomik LC/MS Daten geeignete Programme, welche sowohl frei verfügbar, als auch Open-Source sind: XCMS [Smith et al. 2006] und MZmine [Katajamaa und Oresic 2005]. Beide Programme wurden für die differentielle Analyse von Metabolomik Daten ausgelegt und sind seit dem Jahr 2005 verfügbar.

XCMS

XCMS [Smith et al. 2006] ist Bestandteil von Bioconductor [Gentleman et al. 2004] (`www.bioconductor.org`) einem Open-Source Software Projekt für die Bioinformatik. XCMS kann zur Verarbeitung von sowohl LC/MS als auch GC/MS Daten benutzt werden (daher das X in XCMS). Es sind Importfunktionen für alle gebräuchlichen LC/MS Dateiformate (NetCDF, mzXML, mzData), Funktion zur Feature Detektion, Alignment, Visualisierung sowie einfacher statistischer Analyse der Ergebnisse implementiert. XCMS ist in den Programmiersprachen R und C verfasst, das Paket ist für alle gebräuchlichen Betriebssysteme verfügbar. Die zu diesem Zeitpunkt aktuelle stabile Programmversion ist 1.14.0, vom 21.10.2008.

Der in XCMS ursprünglich implementierte Feature Detektions Algorithmus, im Weiteren *matchedFilter* genannt, beruht auf dem in [Danielsson et al. 2002] beschriebenen Verfahren. Die eingelesenen Rohdaten werden mittels Binning in eine Matrixrepräsentation überführt. Für die chromatographische Analyse der einzelnen Massensignale werden immer zwei benachbarte Zeilen („mass slices“) aus dieser Matrix kombiniert, um das in Abschnitt 3.2.1 beschriebene Problem des Verlustes der charakteristischen Form eines chromatographischen Peaks bei zu klein gewähltem Intervall zu vermeiden. Jedes aus zwei kombinierten „mass slices“ bestehende Chromatogramm wird nun mittels zweifach abgeleiteter Gauss-Funktion gefiltert. Es wird nicht wie bei [Danielsson et al. 2002] oder [Andreev et al. 2003] eine gefilterte LC/MS Matrix erstellt, sondern das gefilterte Chromatogramm wird sofort auf chromatographische Peaks untersucht, die dann gespeichert werden. Dazu wird zunächst aus dem arithmetische Mittel der ungefilterten Intensitätswerte im Chromatogramm ein Schwellwert gebildet. Alle lokalen Maxima der gefilterten Intensitätswerte, die über einem vorgegebenen Vielfachen des Schwellwertes liegen, werden als von einem „echten“ chromatographischen Peak verursacht betrachtet. Ausgehend von den Filter-Maxima werden die jeweiligen Peakgrenzen sowie die Gesamtintensität ermittelt. Die zugehörigen m/z-Werte werden aus der Matrixrepräsentation übernommen und über eine mit den Intensitäten gewichtete Mittelwertbildung als m/z-Koordinate des Features bestimmt. Das fertig bestimmte Feature wird in die Ergebnisliste übernommen.

Dadurch, dass die Kombination der „mass slices“ überlappend vorgenommen wird, d.h. jeder „mass slice“ doppelt verwendet wird, werden Features zum Teil mehrfach detektiert. Ein abschließender Filterschritt eliminiert Features, welche andere Features in bestimmtem Maße überlappen.

MZmine

MZmine [Katajamaa und Oresic 2005] (`mzmine.sourceforge.net`) ist ein in Java geschriebenes und damit plattformunabhängiges Programm zur Visualisierung, Feature Detektion, Alignment und Normalisierung von Metabolomik LC/MS Daten. Es können

NetCDF- und mzXML-Dateien importiert werden. Die zu diesem Zeitpunkt aktuelle stabile Programmversion ist 0.60, vom 09.04.2006. MZmine enthält drei Algorithmen zur Feature Detektion, von denen zwei für die Verarbeitung von Daten im Profilmodus konzipiert sind. Diese beiden Algorithmen bieten zwei verschiedene Varianten für die Centroidisierung der Spektren als Vorverarbeitungsschritt, die Feature-Detektion beruht jedoch bei allen drei Algorithmen auf demselben Verfahren. Für die hier verwendeten centroidisierten Daten ist nur der *Centroid peak detector* geeignet, bei dem ein solcher Vorverarbeitungsschritt entfällt. Dieser Algorithmus wurde für den in Abschnitt 3.4 durchgeführten Vergleich verwendet.
MZmine erstellt zunächst eine Matrixrepräsentation der LC/MS Daten unter Verwendung der Binning-Methode. Die Erkennung einzelner Features findet jedoch, anders als bei XCMS, nicht direkt auf der chromatographischen Ebene statt. Für jedes Chromatogramm, d.h. jeden m/z-bin in der Matrix, wird zunächst ein Schwellwert berechnet, in dem ein vorgegebenes Quantil der Verteilung der Intensitäten berechnet wird (z.B. den Intensitätswert zum 80%-Quantil). Nun werden nacheinander alle Spektren durchlaufen. Für jedes Spektrum werden alle diejenigen Centroide betrachtet, die den jeweils zugehörigen, vorher berechneten chromatographischen Schwellwert überschreiten. Diese Centroide werden über die einzelnen Spektren hinweg miteinander verbunden. Werden keine weiteren Centroide mehr zu einer solchen Kette von Centroiden hinzugefügt, werden Grenzen und Mittelpunkte in RT und m/z bestimmt, die integrierte Intensität ermittelt und das Feature in die Ergebnisliste übernommen. Eine genauere Analyse in chromatographischer Richtung, d.h. bezüglich der chromatographischen Peakform, erfolgt nicht.

3.2.4 Defizite bekannter Algorithmen

Alle bisher bekannten und beschriebenen Verfahren zur Feature-Detektion bei LC/MS Daten haben bestimmte Nachteile, sind schwer zu parametrisieren oder führen nicht immer zum gewünschten Ergebnis. So ist beim vielfach verwendeten Binning wie schon in 3.2.1 beschrieben die Intervallbreite ein kritischer Parameter, der bei Fehleinschätzung verschiedene Probleme verursachen kann und daher sorgfältig optimiert werden muss.
Der in MZmine implementierte Algorithmus findet zwar nach Parameteroptimierung viele Features, liefert aber auch viele falsch positive Features, welche bei genauerer Betrachtung keinen chromatographischen Peak enthalten, sondern lediglich Rauschen. Dies dürfte an der fehlenden Analyse in chromatographischer Richtung liegen, erfolgt die Betrachtung der Feature-Intensitätswerte doch lediglich über eine Schwellwertbetrachtung als Heuristik.
Neben den erwähnten Problemen bei der Anwendung des angepassten Filters, besitzt der *matchedFilter*-Algorithmus aus XCMS noch weitere Nachteile. Zum einen erzeugt ein einzelnes Spike-Signal bei Anwendung des angepassten Filters ein Signal mit derselben Form wie die der Filterfunktion, ohne weitere Überprüfung führt dies zur Erzeugung von falsch

positiven Features. Weiterhin begrenzt der über den einfachen Mittelwert aller Intensitäten in einem Chromatogramm gebildete Schwellwert die Sensitivität des Algorithmus, da kleinere chromatographische Peaks oft nicht erkannt werden, wenn gleichzeitig noch deutlich größere Peaks in diesem Chromatogramm vorhanden sind oder aber eine erhöhte Basislinie vorliegt.
Ein weiterer Algorithmus, konzipiert für die Feature-Detektion auf Proteomik LC/MS Daten [Gröpl et al. 2005] und implementiert in den frei verfügbaren OpenMS-TOPP-Tools [Sturm et al. 2008] (`www.openms.de`) liefert auf Metabolomik Daten derzeit nur sehr unbefriedigende Ergebnisse. Dies lässt sich damit erklären, dass einige Unterschiede zwischen Proteomik und Metabolomik LC/MS Daten existieren (siehe Abschnitt 5.2) und der Algorithmus nur mit und für Proteomik Daten entwickelt wurde. Metabolite werden im Proteomik-Bereich auch eher als „Kontamination“ betrachtet, deren Detektion als Feature nicht erwünscht ist [Trieglaff et al. 2008].

3.3 Der centWave Algorithmus

Aus all diesen Gründen wurde ein verbessertes Verfahren zur Feature-Detektion von Metabolomik LC/MS Daten benötigt. Im Folgenden wird die Technik des neu entwickelten centWave Algorithmus beschrieben. Die centWave Methode besteht im Wesentlichen aus zwei Phasen. In der ersten Phase werden potentiell interessante Regionen von Massensignalen (als *region of interest* (ROI) bezeichnet) detektiert, während in der zweiten Phase eine genaue Analyse aus chromatographischer Sicht, d.h. der Signalintensitäten, unter Benutzung der kontinuierlichen Wavelet Transformation erfolgt.

3.3.1 Phase 1 : Erkennung von Massensignalen

Wie in 3.2.1 diskutiert, ist das Binning zur Separation von Massensignalen ein häufig verwendetes Verfahren, das jedoch einige gravierende Nachteile besitzt. Die beschriebenen Alternativen zum Binning [Stolt et al. 2006, Aberg et al. 2008] leisten deutlich mehr, jedoch mit einer für einen Vorverarbeitungsschritt erheblichen Laufzeit ($\sim$ 5 min bzw. 2 h je Messung[4]). Mit dem hier beschriebenen Verfahren zur Erkennung von Massensignalen wird eine weitere Alternative zum Binning vorgestellt, die in der Lage ist, die in einem LC/MS Datensatz auftretenden Massensignale exakt zu separieren – mit einer deutlich verbesserten Laufzeit ($\sim$ 10-20 Sekunden).
Die für den Algorithmus grundlegenden Beobachtungen (ähnlich wie in [Stolt et al. 2006] beschrieben) sind:

[4]Da die Autoren keine Angaben zur Rechnerkonfiguration machten, sind die Laufzeiten nicht direkt vergleichbar.

- Eine eluierende Substanz, bzw. die daraus gebildeten Ionen zeigen in der m/z-RT-Ebene einer LC/MS Messung im Centroidmodus eine charakteristische Form (siehe dazu Abbildungen 3.5 und 3.6).
- Die m/z-Abweichungen der aufeinanderfolgenden Centroide sind in diesem Bereich gering. Die Höhe der Abweichung ist durch die Genauigkeit des Massenspektrometers bestimmt. Im Bereich der höchsten Peakintensität sind diese Abweichungen am geringsten und steigen dann mit fallender Peakintensität wieder an.
- Die Anzahl der Centroide mit minimaler Abweichung wird durch die chromatographische Peakbreite beschrieben.
- Der in m/z-Richtung direkt benachbarte Bereich neben einem solchen Massensignal ist im allgemeinen leer oder nur dünn besetzt.

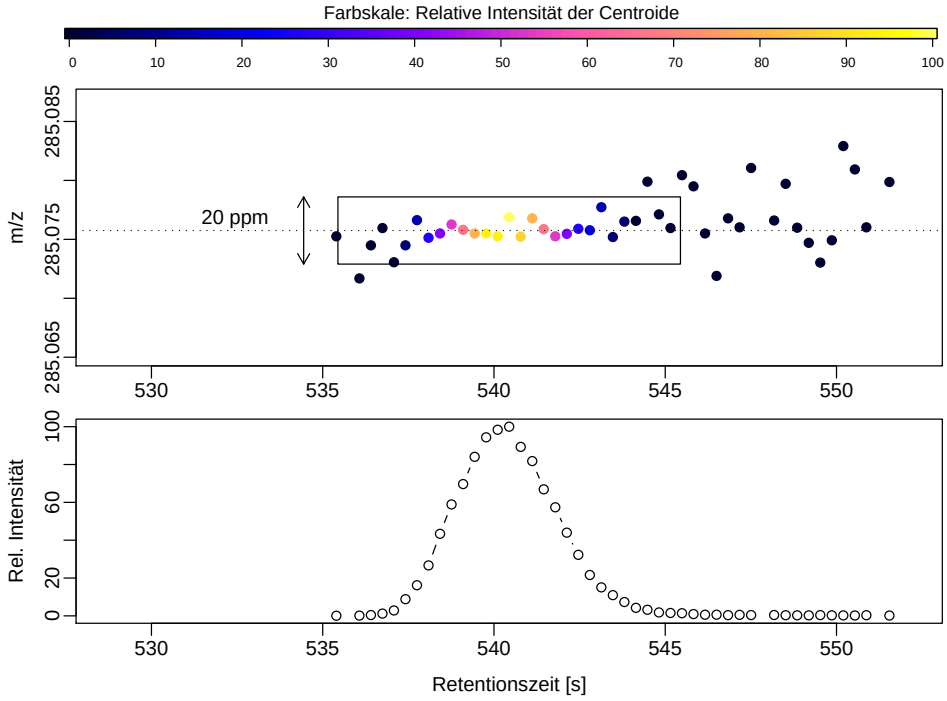

Abbildung 3.5: Die obere Abbildung zeigt das Massensignal des $[M+H]^+$ Ions von Biochanin A über zehn Sekunden. Die Intensitäten sind oben farbkodiert, in der unteren Abbildung als Chromatogramm dargestellt.

Aufgrund dieser Beobachtungen lässt sich vermuten, dass jede interessante Teilregion eines Massensignals – die in der Regel einem chromatographischen Peak entspricht – in der m/z-RT-Ebene durch eine rechteckige Region erfasst werden kann, wobei die Ausdehnung in m/z proportional zur Genauigkeit des Massenspektrometers ist und die Ausdehnung in

der Retentionszeit durch die chromatographische Peakbreite bestimmt ist. Abbildung 3.5 zeigt die *m/z*-RT-Ebene in der Region um ein typisches Massensignal sowie den entsprechenden chromatographischen Peak. Die Genauigkeit des Massenspektrometers, das für diese Messung verwendet wurde beträgt ca. 5 ppm (Herstellerangabe, Bruker MicrOTOF-Q). In der Abbildung ist zu sehen, dass sich die Centroide im Bereich des chromatographischen Peaks maximal ±10 ppm vom theoretischen Wert (gepunktete Linie) entfernen. Die naheliegende Idee für den im Folgenden beschriebenen Algorithmus ist nun, direkt nach solchen *Regionen von Interesse (regions of interest, ROI)* zu suchen, die solche „dichten" Regionen bilden. Dazu sind mindestens zwei Parameter erforderlich, die minimale Ausdehnung einer solchen ROI in zeitlicher Richtung und die erlaubte Abweichung der Centroide in *m/z*-Richtung. Diese beiden Parameter sind leicht einzuschätzen, da die auftretende minimale chromatographische Peakbreite und die Gerätegenauigkeit dem Experimentator bekannt sind.
Der Grundprinzip lässt sich wie folgt beschreiben: Die massenspektrometrischen Scans der Messung werden in ihrer zeitlichen Abfolge betrachtet. Dabei werden ROI's, welche eine Bereich zeitlich aufeinanderfolgender Centroide beschreiben, inkrementell erweitert bzw. neue ROI erzeugt.
Zur Initialisierung wird zunächst jeder Centroid des ersten Scans als neue ROI aufgenommen. Für jeden folgenden Scan wird nun überprüft, welche Centroide darin innerhalb der vorgegebenen Abweichung zu den in *ROI().mzmean* gespeicherten *m/z*-Mittelwerten jeder ROI liegen. Diese Centroide werden an die jeweils entsprechende ROI angehangen, und der Eintrag in *ROI(j).mzmean* als Mittelwert der in der *j*-ten ROI gespeicherten *m/z*-Werte aktualisiert. Für neu auftretende Centroide werden neue ROI begonnen. Die Aufnahme von Centroiden in eine ROI wird abgebrochen, wenn die Abweichung die vorgegebene Toleranz überschreitet. Nach jedem Scan werden die aktuell bearbeiteten ROI überprüft. Falls aus dem aktuellen Scan kein Centroid zugefügt wurde und die ROI weniger als die vorgegebene minimale Anzahl von Centroiden enthält, wird sie gelöscht. Enthält sie genügend Centroide und „wächst" nicht mehr, gilt die ROI als abgeschlossen und wird in die Ergebnisliste übertragen.
Der in jedem Schritt aktualisierte Wert *ROI(j).mzmean* einer Region *j* markiert somit den momentan geschätzten *m/z*-Mittelpunkt und dient als Vergleichswert im nächsten Schritt der Iteration. Die *m/z*-Ausdehnung der ROI ist folglich nicht starr, sondern folgt in einem gewissen Maß dem Verlauf der Centroide eines Massensignals. Dies ist notwendig, um detektorabhängige Effekte (siehe Abbildung 3.14 Seite 45 und Beschreibung dort) auszugleichen, die bestimmte Schwankungen im Verlauf der Centroide hervorrufen.
Der in Listing 1 in Pseudocode beschriebene Algorithmus zur Detektion von den ROI in einer LC/MS Messung benötigt die folgenden Eingabeparameter:

- M : LC/MS Messung im Centroid-Modus, mit Scan-Nummern $1, \ldots, S$

- μ : Tolerierte Massenabweichung, gegeben in ppm

- $\mathrm{p_{min}}$: Mindestbreite eines chromatographischen Peaks in Scans, diese wird unter Benutzung der Scan-Rate der Messung berechnet aus der Nutzereingabe $\mathrm{peakwidth}_{min}$, der Mindestbreite eines chromatographischen Peaks in Sekunden.

Optional können noch die Parameter k und I übergeben werden, um Regionen mit sehr geringer Intensität zu filtern. Dieser Schritt verkürzt die Laufzeit der zweiten Phase der Feature-Detektion, der chromatographischen Analyse aller ROI (Abschnitt 3.3.2).
Der für die ROI-Detektion benötigte Parameter $\mathrm{p_{min}}$ wird über die untere Grenze des berechneten Skalenbereichs (Abschnitt 3.3.2) abgeschätzt: $\mathrm{p_{min}} = \max(5,\mathrm{MinScale})$.
Der Parameter $\mathrm{p_{min}}$ wird bewusst etwas kleiner als die minimale Skale gewählt, damit sichergestellt ist, dass bei der ROI-Detektion ausreichend viele Centroide innerhalb des Fenster μ liegen und an dieser Stelle keine evtl. interessanten Regionen übersehen werden.
Abbildung 3.6 visualisiert das Ergebnis der ROI-Detektion für einen kleinen Ausschnitt aus einer LC/MS Messung.

Betrachtung der Laufzeit

Jeder Centroid mz_i^s wird genau einmal betrachtet und mittels binärer Suche mit dem jeweils aktuellen *ROI().mzmean*-Vektor verglichen (Zeile 18). Sei die Anzahl der Centroide $C = |M|$, dann lässt sich die Anzahl $\overline{J}$ der zu einem Zeitpunkt bearbeiteten ROI über die mittlere Anzahl von Centroiden pro Scan abschätzen: $\overline{J} = \frac{C}{S}$. Die mittlere Anzahl durchgeführter Vergleiche beträgt somit $C \cdot log_2\overline{J}$. Wird eine ROI erweitert, so muss der *m/z*-Mittelwert der enthaltenen Centroide aktualisiert werden (Zeile 22). Da dies unter Verwendung des bisherigen Mittelwerts geschehen kann, hat dieser Schritt nur eine Laufzeit von $\mathcal{O}(1)$. Bei der Erstellung einer neuen ROI muss auch ein neuer *m/z*-Wert in den „Verzeichnis“-Vektor *ROI().mzmean* eingetragen werden (Zeile 25, Funktion *addROI*)[5]. Da dieser stets sortiert gehalten wird, kann auch das Suchen der Einfügestelle mittels Binärsuche erfolgen (Laufzeit $log_2\overline{J}$). Es kann maximal jeder Centroid eine neue ROI erzeugen, also beträgt auch die Laufzeit dieses Schrittes insgesamt $C \cdot log_2\overline{J}$.
Der Validierungsschritt (Zeile 28) betrachtet alle gerade bearbeiteten ROI. Es werden solche ROI gelöscht, die in der Iteration s nicht erweitert wurden und weniger als $\mathrm{p_{min}}$ Centroide enthalten. Wurde eine ROI i nicht mehr erweitert, enthält aber mindestens $\mathrm{p_{min}}$ Centroide, von denen k eine Intensität $\geq I$ besitzen ($\mathrm{ROI}(i).\mathrm{intCount} \geq \mathrm{k}$), so wird diese ROI in die Ergebnisliste übertragen und aus der Menge der aktuellen bearbeiteten ROI's entfernt. Jede der gerade bearbeiteten ROI wird pro Scan einmal überprüft. Die Laufzeit des Validierungsschritts beträgt somit $S \cdot \overline{J} = S \cdot \frac{C}{S} = C$. Die Gesamtlaufzeit des

[5]Die Verbindung zwischen dem jeweiligen Mittelwert $\mathrm{ROI}(i).\mathrm{mzmean}$ und den eigentlichen Werten in $\mathrm{ROI}(i).\mathrm{values}$ ist in der Implementierung über Zeiger realisiert.

```
1  Eingabe: M = {m_i^s | 1 ≤ s ≤ S, 1 ≤ i ≤ N_s}, μ, p_min
2           optionale Parameter : k, I // sonst null gesetzt
3  Ausgabe: ROI_final // Liste von ROI
4
5  // Initialisiere ROI mit erstem Scan
6  for all i = 1,...,N_s, N_s = |mz^1| do
7       ROI(i).values(1) = mz_i^1
8       ROI(i).mzmean = mz_i^1
9       if (intensity(mz_i^1) ≥ I)
10      then  ROI(i).intCount = 1
11      else  ROI(i).intCount = 0
12 end for
13 // Verarbeitung aller nachfolgenden Scans
14 for all s = 2,...,S do
15      for all i = 1,...,N_s, N_s = |mz^s| do
16          // Existiert ein j, j = 1,...,J, J = |ROI|, so dass
17          // |ROI(j).mzmean − mz_i^s| <= μ ?
18          j= binarySearch(ROI(1,...,J).mzmean, mz_i^s, μ)
19          if (j >= 1)
20          then  {// Hänge mz_i^s an ROI(j) an
21              K = |ROI(j).values| + 1 , ROI(j).values(K) = mz_i^s
22              updateMean(ROI(j).mzmean, K, mz_i^s)
23              if (intensity(mz_i^s) ≥ I)
24              then  ROI(j).intCount += 1 }
25          else addROI(ROI, mz_i^s) // Erstelle neue ROI
26      end for
27      // Überprüfen, evtl. Übernehmen der aktuellen ROI
28      ROI_final += validate(ROI, p_min, k)
29 end for
```

Listing 1: Algorithmus zur ROI Detektion in LC/MS Daten

ROI-Algorithmus beträgt also $2 \cdot C \cdot log_2 \overline{J} + C$ und liegt damit in der Komplexitätsklasse $\mathcal{O}(C \cdot logC)$.

Die in dieser Arbeit verwendeten LC/MS Messungen enthalten zwischen 2500 und 3000 Scans und ca. 2.5 bis 4.5 Millionen Centroide. Die Laufzeit des ROI-Algorithmus auf derartigen Messungen beträgt $\approx$ 10-20 Sekunden auf einer 2.5 GHz CPU.

In einer auch im nächsten Abschnitt als Beispiel verwendeten LC/MS-Messung (HPLC an

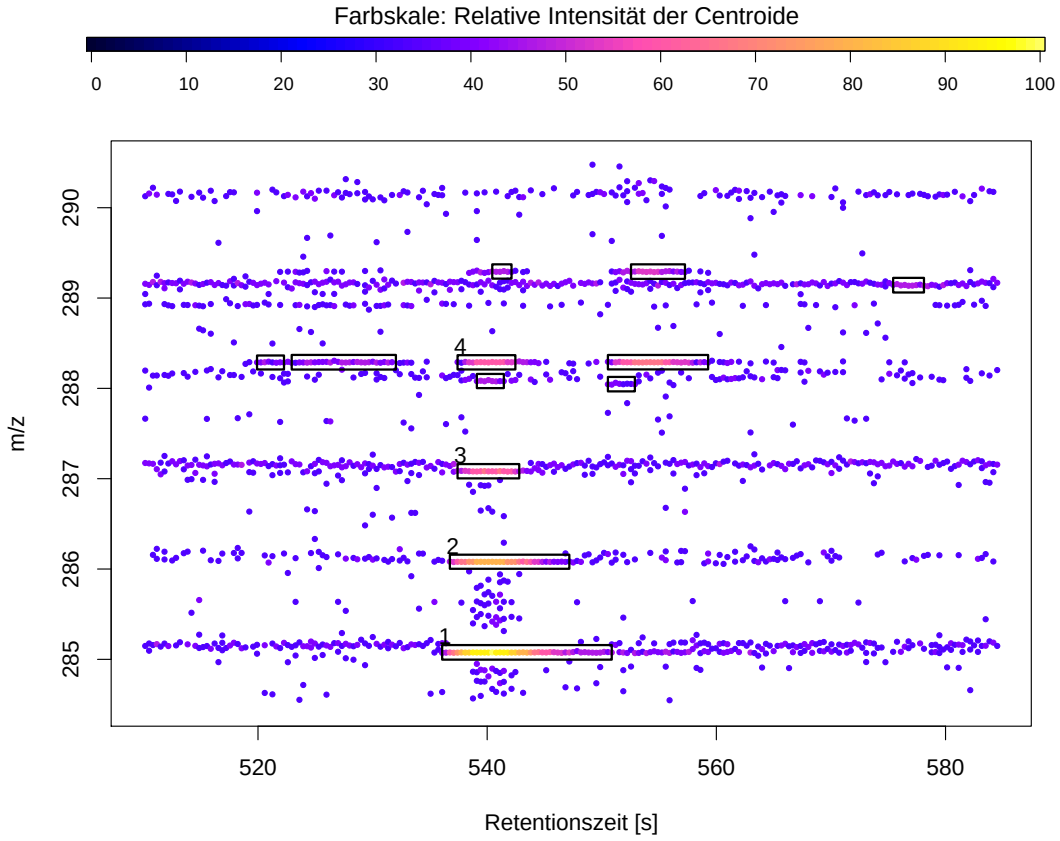

Abbildung 3.6: LC/MS Rohdaten und detektierte ROI's in der Region um das $[M+H]^+$ Massensignal (1) von Biochanin A. Die Regionen (2-4) sind die zu (1) gehörigen Isotopomere, die anderen Regionen sind unbekannte Signale. Zur besseren Visualisierung wurde die m/z-Ausdehnung jeder ROI um 0.15 m/z erhöht.

einem QTOF-Massenspektrometer, Datensatz E105) wurden mit den Parametern μ=10 ppm, p_{min}=5, k=3, I=300 insgesamt 5699 ROI gefunden. Die durchschnittliche Länge einer ROI, angegeben in Scans, betrug dabei 8.8 (Median 7). Abbildung 3.7 zeigt die Verteilung der Länge für die erfassten 5699 ROI.

Im Vergleich zum Binning-Verfahren besitzt der ROI-Ansatz den Vorteil, dass keine feste Intervallbreite gewählt werden muss. Jede ROI wird separat detektiert und die Nachteile des Binning-Verfahrens können so vermieden werden. Im Gegensatz zum Binning ist das Ergebnis jedoch keine Matrix, sondern eine Liste von potentiell interessanten Regionen. Abhängig von der chromatographischen Trennung, der Genauigkeit des Massenspektrometers, der Komplexität der Probe und dementsprechend der Featuredichte in der Messung, kann jede ROI keinen, genau einen oder mehrere ausgeprägte chromatographische Peaks enthalten. Deshalb ist es notwendig, jede ROI einer umfangreichen Analyse in der chromatographischen Domäne zu unterziehen.

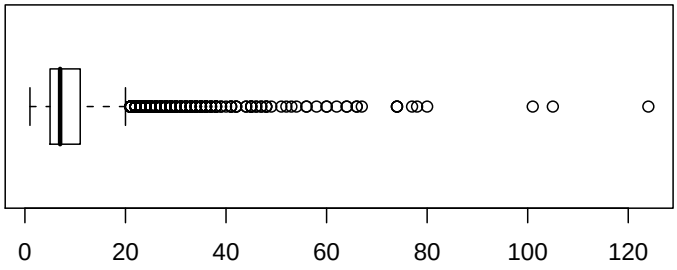

Abbildung 3.7: Boxplot für die Länge von 5699 ROI aus einer LC/MS Messung, angegeben in Scans.

3.3.2 Phase 2 : Feature-Detektion

Wavelet-basierte Methoden wurden im Bereich der Massenspektrometrie bereits für die Peak Detektion in einzelnen Massenspektren von Proteomik Daten verwendet [Lange et al. 2006, Du et al. 2006, Conrad et al. 2006, McLerran et al. 2008]. Im Folgenden wird die Anwendung der kontinuierlichen Wavelet Transformation für die Erkennung von chromatographischen Peaks bei Metabolomik LC/MS Daten gezeigt.

Die kontinuierliche Wavelet Transformation
Die klassische Fouriertransformation, definiert als

$$F(\omega) = \int_{-\infty}^{\infty} f(t)\mathrm{e}^{-\mathrm{i}\omega t}\mathrm{d}t \quad \text{und} \quad f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)\mathrm{e}^{\mathrm{i}\omega t}\mathrm{d}\omega$$

liefert Information über den Frequenzgehalt des *gesamten* Signals $f(t)$. Eine Aussage bezüglich der Zeitpunkte, an denen eine bestimmte Frequenz ω im Signal auftritt, ist damit nicht möglich.
Die Kurzzeit-Fouriertransformation (short time fourier transformation, STFT) ermöglicht eben solche Aussagen, in dem eine (fest gewählte) Fensterfunktion $g(t)$ eingeführt wird :

$$STFT(\omega, \tau) = \int_{-\infty}^{\infty} f(t)g(t-\tau)\mathrm{e}^{-\mathrm{i}\omega t}\mathrm{d}t$$

Das Ergebnis beschreibt die Frequenzanteile ω von $f(t)$ in der Umgebung des Zeitpunktes τ. Wählt man das Fenster

$$g(t) = \frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{t^2}{2\sigma^2}}$$

mit $\sigma > 0$ als festem Parameter, so erhält man eine verbreitete Variante der Kurzzeit-Fouriertransformation, die *Gabor-Transformation* (nach dem ungarischen Mathematiker

Dénes Gábor). Die Kurzzeit-Fouriertransformation besitzt den Nachteil, dass durch die feste Wahl der Fensterfunktion g sowohl die Auflösung in der Zeit als auch in der Frequenz festgelegt wird. Wird eine schmale Fensterbreite gewählt, können nur hohe Frequenzen optimal zeitlich aufgelöst werden, während das Fenster dann aber zu schmal ist, um volle Amplituden einer niedrigen Frequenz zu erfassen. Wird andererseits eine breite Fensterfunktion gewählt, können dann zwar niedrige Frequenzen genau erfasst werden, jedoch wird dies mit einer schlechteren zeitlichen Auflösung der hohen Frequenzen erkauft. Für die Praxis wünschenswert wäre jedoch eine hohe zeitliche Auflösung für hohe Frequenzanteile bei gleichzeitiger Erfassung niedriger Frequenzen, die sich über ein längeres Zeitfenster erstrecken.
Diese Anforderungen erfüllt die kontinuierliche Wavelet Transformation (Continuous Wavelet Transform, CWT). Die Fouriertransformation verwendet als analysierende Funktion $e^{\mathrm{i}t}$, skaliert mit dem reellen Frequenzparameter α: $e^{\mathrm{i}\alpha t}$. Die Kurzzeit-Fouriertransformation verwendete dieselbe analysierende Funktion $e^{\mathrm{i}\alpha t}$ in Verbindung mit einer starren Fensterfunktion g. Im Unterschied zur Fouriertransformation verwendet die Wavelet Transformation anstelle der Frequenz die sogenannte Skalierung, welche sich umgekehrt proportional zur Frequenz verhält. Als analysierende Funktion wird eine Wavelet-Funktion ψ verwendet, welche mit den Parametern s und τ skaliert bzw. verschoben wird:

$$\psi_{s,\tau}(t) = \frac{1}{\sqrt{s}}\psi\left(\frac{t-\tau}{s}\right),\ s \in \mathbb{R}^+ - \{0\},\ \tau \in \mathbb{R} \tag{3.4}$$

Ein Beispiel für eine häufig gewählte Wavelet-Funktion ist die auf $||\psi|| = 1$ normierte negierte zweite Ableitung der Gauss-Funktion, aufgrund seiner Form auch als „Mexican Hat“ bekannt:

$$\psi(t) = \frac{2}{\sqrt{3}}\pi^{-\frac{1}{4}}\left(1-t^2\right)e^{\frac{-t^2}{2}} \tag{3.5}$$

Abbildung 3.8 zeigt das Mexican Hat Wavelet bei verschieden Skalierungen s.
Sei $\psi(t)$ eine reellwertige Funktion, dann ist die Wavelet-Transformierte des Signals $f(t)$ wie folgt definiert:

$$W_f(s,\tau) = \int_{-\infty}^{\infty} f(t)\psi_{s,\tau}(t)\mathrm{d}t \tag{3.6}$$

und liefert für diskrete Werte von s, τ die Matrix W_f der Waveletkoeffizienten. Ähnlich wie bei der STFT beschreibt W_f Frequenzanteile von $f(t)$ in der Umgebung des Zeitpunktes τ, welche jedoch über die Skale s (statt der Frequenz ω) beschrieben werden. Der wichtigste Unterschied aber ist der durch die Verwendung der Skalierung erzielte Vorteil, dass die Breite des Abfragefensters *nicht* (wie bei der STFT) konstant ist. Wie in Abbildung 3.8 sichtbar, wächst die Breite des Abfragefenster proportional zum Betrag der Skale s. Hohe Skalenwerte ergeben ein breites Fenster und sind geeignet, um langwellige Schwingungsanteile zu erfassen. Niedrige Skalenwerte hingegen ermöglichen den präzise lokalisierten

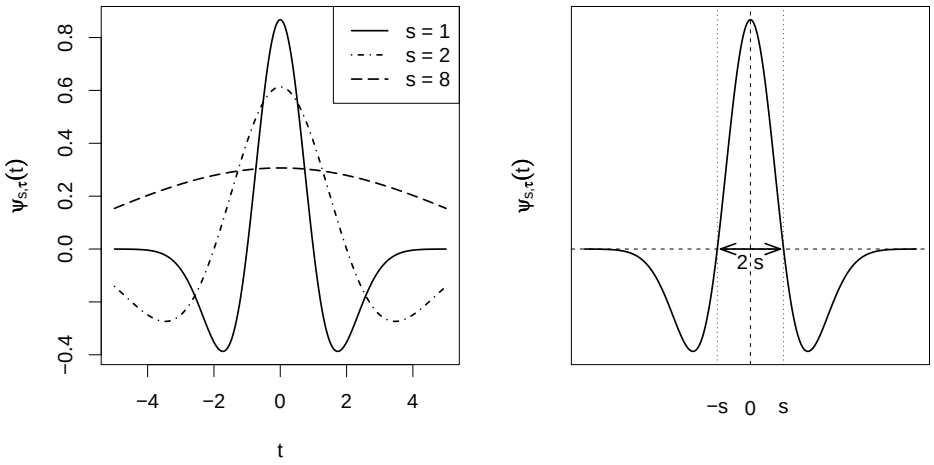

Abbildung 3.8: Das Mexican Hat Wavelet bei verschiedenen Skalierungen (links), Nullstellen und Basisbreite bei Skalierung s (rechts).

Nachweis von hohen Frequenzanteilen [Daubechies 1992, Blatter 2003, Meffert und Hochmuth 2004].

Anwendung der CWT zur Detektion chromatographischer Peaks

Die negierte, zweifach abgeleitete Gauss-Funktion hat sich als angepasster Filter zur Detektion chromatographischer Peaks bewährt. Verwendet man das Mexican Hat Wavelet mit nur einer – entsprechend gewählten – Skala, so ist das Ergebnis vergleichbar mit der angepassten Filterung mittels zweifach abgeleiteter Gauss-Funktion. Das gefilterte Signal unterscheidet sich vom Ergebnis der Wavelet-Transformation dann nur durch einen Faktor.

Um die in Abschnitt 3.2.2 beschriebene Probleme bei der Verwendung des angepassten Filters zu vermeiden – insbesondere die schwierige Entscheidung für die „richtige“ Breite des Modellpeaks – bietet sich die Untersuchung der Chromatogramme auf *mehreren* Skalen mittels der kontinuierlichen Wavelet Transformation als Lösung an.

Der Benutzer soll nun nicht mehr eine einzige, „richtige“ Breite der chromatographischen Peaks vorgegeben, sondern einen Größenbereich, in dem er die Peakbreite vermutet bzw. durch vorheriger Experimente kennt. Die Praxis zeigt, dass hierfür wirklich nur eine ungefähre Einschätzung nötig ist. Aus dem gegebenen Minimum und Maximum der Peakbreite können die Skalen festgelegt werden, auf denen die CWT berechnet wird.

Setzt man das „Mutter“-Wavelet 3.5 in die Gleichung 3.4 ein und setzt den Verschiebungsparameter $\tau = 0$, so ergibt sich:

$$\psi_s(t) = \frac{1}{\sqrt{s}} \frac{2}{\sqrt{3}} \pi^{-\frac{1}{4}} \left(1 - \left(\frac{t}{s}\right)^2\right) e^{\frac{-\left(\frac{t}{s}\right)^2}{2}} \tag{3.7}$$

$\psi_s(t)$ besitzt die Nullstellen $t = -s$ und $t = s$, die Basisbreite des Mexican Hat Wavelets bei der Skalierung s beträgt $2s$ (Abbildung 3.8). Die verwendete Skale, bzw. die Breite des skalierten Wavelets, sollte in ungefähr der Breite des zu detektierenden Signals entsprechen. Aus praktischen Gründen werden die chromatographischen Peaks auf einer Zeitachse betrachtet, die nicht in Sekunden sondern, mit Scan-Nummern beschriftet ist. Da die Scanrate konstant ist und somit die einzelnen Scans den gleichen Abstand haben, macht dies bezüglich der Form der Peaks keinen Unterschied, erleichtert aber die Verarbeitung, da die Scan-Nummer leichter zu adressieren ist.

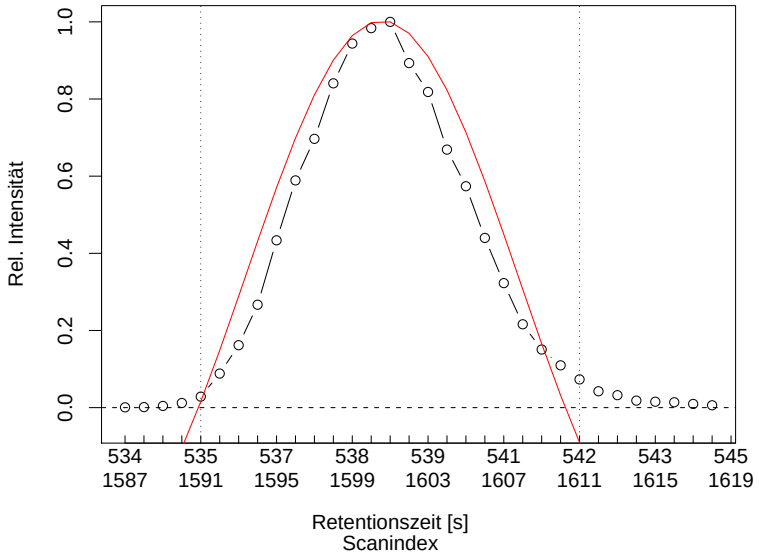

Abbildung 3.9: Beispiel für die Skale eines chromatographischen Peaks. Die Breite (senkrechte Striche) bei diesem Peak beträgt ca. 7 s, bei einer Scanrate von 3 Hz. Rot darübergelegt ist der pos. Teil eines Mexican Hat Wavelets mit der Skale $s = 10$.

Sei ISD der Abstand zwischen zwei Scans (inter-scan-distance) in Sekunden, so kann nun aus der Benutzereingabe peakwidth_{min} und peakwidth_{max} (geschätzte Mindest- und Höchstbreite der auftretenden chromatographischen Peaks in Sekunden) der entsprechende Skalenbereich abgeleitet werden:

$$\begin{aligned} \text{MinScale} &= round((\text{peakwidth}_{min}/\text{ISD})/2) \\ \text{MaxScale} &= round((\text{peakwidth}_{max}/\text{ISD})/2) \end{aligned}$$

Der Skalenbereich wird sowohl für die Wavelet-Transformation also auch für die Bestimmung anderer Parameter benötigt. Abbildung 3.9 zeigt ein Beispiel für die Skale eines chromatographischen Peaks.

Beschreibung der Verarbeitungsschritte
Im Folgenden werden die einzelnen Verarbeitungsschritte der zweiten Phase (Feature-Detektion) des centWave Algorithmus aufgeführt. Listing 2 zeigt den Ablauf im Überblick. Benötigt werden die folgenden Eingabeparameter

- ROI : die in Phase 1 detektierten *Regions Of Interest* (siehe Listing 1)
- peakwidth_{min}, peakwidth_{max} : geschätzte Mindest- und Höchstbreite der auftretenden chromatographischen Peaks in Sekunden
- SNR_{Thr} : Schwellwert für das Signal/Rausch-Verhältnis
- fitGaussian : Angabe, ob an jeden chromatographischen Peak eine Gausskurve angepasst werden soll.

1. Schätzung von Basislinie und Rauschpegel
Die Massenabweichung der aufeinanderfolgenden Centroide ist normalerweise bei hoher Intensität der Centroide am geringsten (siehe auch Punkt 5 auf Seite 44). Mit sinkender Intensität vergrößert sich die Massenabweichung. Unter Umständen wird daher bei der ROI-Detektion, vor allem bei geringer Intensität des Features und abhängig von der tolerierten Massenabweichung μ, nur der Kernbereich eines Features detektiert. Um in jedem Fall den ganzen chromatographischen Peak zu erfassen und außerdem einen Bereich zur Schätzung von Basislinie und Rauschpegel bereitzustellen, wird die Ausdehnung der untersuchten Region in RT-Richtung auf das dreifache der vorgegebenen maximalen Peakbreite erweitert.
In dieser erweiterten ROI wird zunächst ein Schätzung der chromatographischen Basislinie vorgenommen. Eine solche Basislinie wird sowohl durch eine Art „Grundrauschen“ vom Detektor verursacht, aber auch durch im gesamten Retentionszeitbereich auftretende Hintergrundsignale wie Lösungsmittel etc. („chemisches Rauschen“). Letztere treten zwar nur in bestimmten Massenbereichen auf, können dann aber eine vergleichsweise hohe Basislinie erzeugen, welche die chromatographischen Peaks von Substanzen mit ähnlicher Masse zum Teil verdeckt. Eine erste Abschätzung dieser Basislinie wird über den 5%-getrimmten Mittelwert der Intensitätswerte in der erweiterten Region vorgenommen.
Eine Teilmenge der detektierten ROI's enthält lediglich eine solche Basislinie, jedoch keinen erkennbaren chromatographischen Peak. Um solche ROI schon vor den weiteren, zeitaufwendigeren Schritten auszufiltern, wird eine einfache Heuristik verwendet. Dazu wird untersucht, ob mindestens n aufeinanderfolgende Intensitätswerte über dieser Basislinie liegen ($n = 2 \cdot \text{MinScale}$). Ein eindeutiger chromatographischer Peak ist dadurch gekennzeichnet, dass er aus Datenpunkten besteht, die sich in ihrer Intensität gegenüber der Basislinie deutlich abheben und deren Intensitätsverlauf in etwa die Form einer Gauss-Kurve

```
1  Eingabe: ROI, peakwidth_min, peakwidth_max, SNR_Thr, fitGaussian
2  Ausgabe: F // Liste von LC/MS Features
3
4  scales=getScales(peakwidth_min, peakwidth_max)
5  for all i = 1,...,N, N = |ROI| do
6      eROI=extendROI(ROI(i))
7      BL=estimateRegionBaseline(eROI)
8      if !continuousPtsAboveBaseline(eROI, BL)
9      then next
10     (NL,BL)=getLocalNoiseEstimate(eROI)
11     wCoefs= CWT(eROI, scales, wavelet='mexHat')
12     localMax=getLocalMaximumCWT(wCoefs)
13     ridgeList=getRidge(localMax)
14     for all j = 1,...,K, K = |ridgeList| do
15         ft=newFeature()
16         (ft.Scale,ft.RTCenter)=getBestScale(ridgeList_j,eROI)
17         ft.RTRange=descendMin(eROI,wCoefs, ft.Scale)
18         (ft.Intensity,ft.MaxInt)=getInt(eROI, ft.RTRange)
19         if (FitGaussian)
20         then {ft.Gauss=fitGauss(eROI,ft.RTRange)
21               ft.RTRange=getOptRTRange(ft.Gauss,ft.RTRange)
22              }
23         ft.SN= (ftMaxInt - BL) / NL)
24         if (ft.SN < SNR_Thr)
25         then next
26         (ft.MZCenter,ft.MZRange)=getMZValues(eROI,ft.RTRange)
27         F= addFeature(F, ft)
28     end for
29 end for
```

Listing 2: Phase 2 des centWave Algorithmus : Feature-Detektion

beschreibt. Aufgrund dieser Charakteristik können ROI, welche die obige Bedingung nicht erfüllen, verworfen werden. Von den in der Beispielmessung detektierten 5699 ROI werden in diesem Schritt 2375 ROI (≈ 42 %) verworfen.

Als nächstes werden Mittelwert und Standardabweichung der Intensitätswerte der erweiterten ROI *ohne* den ursprünglich detektierten Bereich berechnet, sozusagen „links“ und „rechts“ vom mutmaßlichen chromatographischen Peak. Der Mittelwert wird mit der

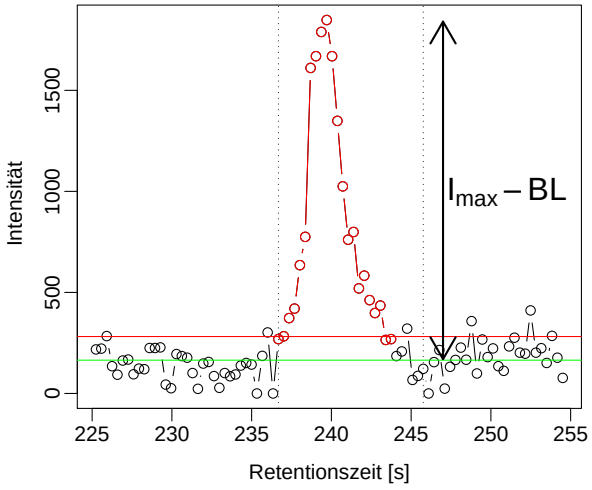

Abbildung 3.10: Heuristische Peak-Vorfilterung und Schätzung der Basislinie. Die gestrichelten Linien markieren den ursprünglich als ROI detektierten Bereich. Die erste Schätzung der Basislinie sowie die zusammenhängend über dieser liegenden Intensitätswerte sind rot markiert. Die finale Basislinie ist grün gezeichnet. In diesem Beispiel ist $\mathrm{SNR} = \frac{\mathrm{I_{max}-BL}}{\mathrm{NL}} = (1848 - 282)/63 = 25$.

ersten Abschätzung der Basislinie kombiniert, indem das Minimum aus beiden gewählt wird. Das Resultat liefert die endgültige Schätzung der Basislinie (BL) dieser ROI. Grund für diese vorsichtige Wahl ist es, in Regionen mit mehreren chromatographischen Peaks kleinere Peaks nicht zu benachteiligen, da durch die Nachbarschaft großer Peaks der Intensitätsmittelwert relativ hoch ausfallen kann und dann die untere Abschätzung häufig die günstigere Wahl darstellt. Die Standardabweichung der gesamten umgebenden Region wird als Maß für den Rauschpegel (NL) in der ROI benutzt.

Da im Bereich der Massenspektrometrie und insbesondere für LC/MS noch keine bzw. keine einheitliche Methoden für die Schätzung von Signal/Rausch-Verhältnis (SNR) von Features bzw. des chromatographischen Peaks existieren, wurde gemeinsam mit einem Chemiker (Christoph Böttcher, IPB Halle) und in Anlehnung an die im Bruker „Target-Analysis“ Benutzerhandbuch [Bruker 2007] angegebene Methode folgende Definition erarbeitet:

$$\mathrm{SNR} = \frac{\mathrm{I_{max}} - \mathrm{BL}}{\mathrm{NL}}$$

$\mathrm{I_{max}}$ bezeichnet den maximale Intensitätswert des chromatographischen Peaks. Die in [Bruker 2007] angegebene Variante ist

$$\mathrm{SNR}_{Bruker} = \frac{\mathrm{I_{max}} - baseline}{5\sigma}$$

wobei σ berechnet wird als die Standardabweichung aller Werte größer eines bestimmten Schwellwerts aus der dritten Ableitung des Chromatogramms, welche wiederum mittels

Savitzky-Golay-Algorithmus durchgeführt wurde. Details dazu sowie die Art der Ermittlung der Basislinie *baseline* sind nicht angegeben.
Da nunmehr Basislinie und Rauschpegel der Region bestimmt sind, wird die aktuelle ROI nur dann weiter betrachtet, wenn sie mindestens einen Intensitätswert I_x enthält, der das vom Benutzer geforderte Signal/Rausch-Verhältnis SNR_{Thr} erfüllt, d.h. für den gilt:

$$I_x - BL \geq SNR_{Thr} * NL$$

2. Berechnung der kontinuierlichen Wavelet Transformation
Als nächster Schritt wird für die erweiterte Region die kontinuierliche Wavelet Transformation mit dem Mexican Hat Wavelet auf den Skalen (MinScale,. . .,MaxScale) berechnet. Abbildung 3.11 zeigt das Ergebnis der CWT für eine Region, die einen hohen chromato-

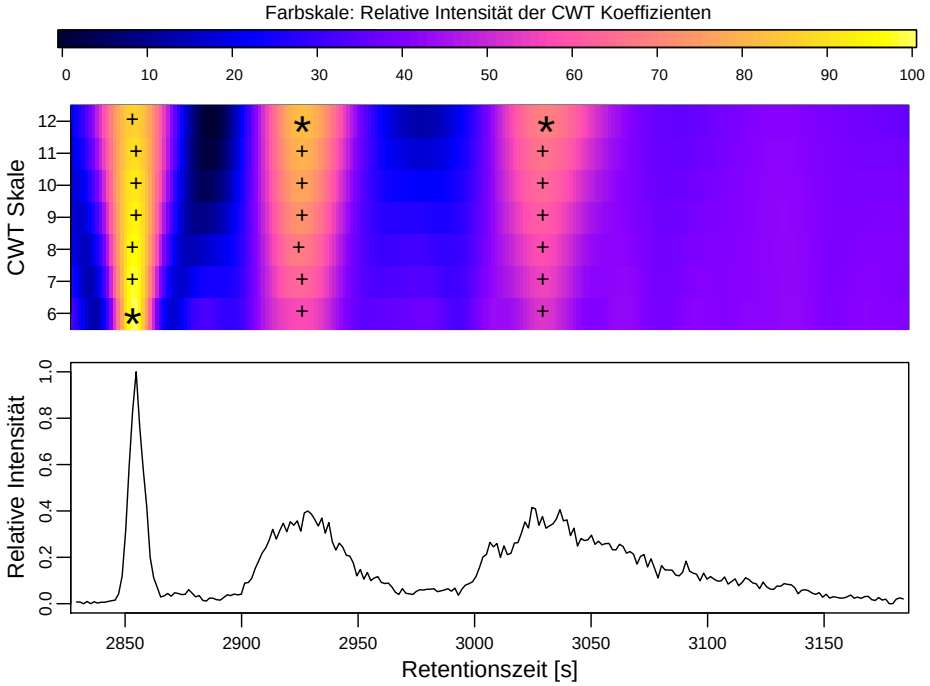

Abbildung 3.11: Berechnete CWT Koeffizienten für ein Chromatogramm mit drei Peaks. Die lokalen Maxima der Koeffizienten wurden zu *Ridge Lines* verbunden. Der als optimale Beschreibung für den entsprechenden Peak ausgewählte Punkt der Ridge Line ist mit einem Stern markiert.

graphischen Peak geringer Breite, sowie zwei schwächer ausgeprägte Peaks größerer Breite enthält. Verwendet wurden die Skalen 6 bis 12. Zu erkennen sind die unterschiedlichen Intensitäten der CWT-Koeffizienten auf der jeweiligen Skale. Die der Region um den hohen, schmalen Peak (bei ca. 2855 s) entsprechenden Koeffizienten zeigen die höchste Intensität auf der niedrigsten Skale, während die beiden breiteren Peaks (bei ca 2925 und 3030 s) hohe Koeffizientenwerte auf den höheren Skalen ergeben.

3. Erkennen und Verbinden der lokalen Maxima
Um die einzelnen Peaks auf den jeweiligen Skalen zu lokalisieren, werden lokale Maxima auf jeder Skale gesucht und zu sogenannten *Ridge Lines* verbunden. Für jede Skale i werden zunächst mittels eines gleitenden Fensters, dessen Breite proportional zur Skale gewählt wird ($w_i = scales_i * 2 + 1$), lokale Maxima detektiert. Um aus den Maxima auf den jeweiligen Skalen die Ridge Lines zu bilden, werden die Ridge Lines zunächst initialisiert mit den lokalen Maxima der größten Skale. Als nächstes werden die in der nächstkleineren Skale i lokalisierten Maxima untersucht. Wenn sich innerhalb des Fenster w_i ein Maximum auf dieser Skale befindet, so wird dieses zur jeweiligen Ridge Line hinzugefügt. Befinden sich mehrere Maxima innerhalb dieses Fensters, so wird nur das mit dem kleinsten Abstand zur Position des Maximums in der aktuellen Skale gewählt. Diese Prozess wird bis zur kleinsten Skale fortgeführt, wobei erlaubt ist, dass Ridge Lines enden, wenn kein weiteres Maximum in der nächstkleineren Skale vorliegt, oder aber bei neuen, nach oben nicht verbundenen Maxima, neu anfangen. Das Ergebnis dieses Schrittes für das Beispiel ist ebenfalls in Abbildung 3.11 gezeigt. Die lokalen Maxima, die eine Ridge Line bilden, sind dabei markiert. Der Algorithmus für diesen Schritt wurde weitgehend von [Du et al. 2006] übernommen, der in dem R-Paket *MassSpecWavelet*[6] implementiert ist.

4. Bestimmung der chromatographischen Parameter
Ausgehend von den Ridge Lines können nun chromatographische Peaks lokalisiert werden. Die Ridge Line beschreibt den Skalenbereich, auf dem ein Peak detektiert wurde. Die Skale, auf der die Waveletkoeffizienten den maximalen Wert haben, beschreibt den Peak theoretischerweise optimal, d.h. der vorliegende chromatographische Peak sollte in seiner Breite in etwa der doppelten Skale entsprechen. Dies ist jedoch dann nicht der Fall, wenn eng benachbarte, oder überlappende chromatographische Peaks auftreten (siehe Abschnitt 3.3.3). In diesen Fällen haben z.T. größere Skalen, die die benachbarten bzw. überlappenden Peaks als Ganzes beschreiben, höhere Koeffizienten. Die Betrachtung allein der Skale mit den höchsten Koeffizienten ist somit nicht ausreichend. Es wird daher für jeden Punkt der Ridge Line, d.h. jedes der darin enthaltenen lokalen Maxima der Waveletkoeffizienten, überprüft, welche Intensitätswerte an der entsprechenden Position im Chromatogramm vorliegen. Um eine gewisse Toleranz gegenüber lokalen Schwankungen der Intensitätswerte zu gewährleisten, wird nicht nur ein einzelner Wert betrachtet, sondern über ein Fenster mit der Breite der minimalen Skale summiert. Aus diesen Intensitätswerten wird das Maximum ausgewählt. Falls mehrere Werte gleich hoch sind, wird zusätzlich die Höhe der Waveletkoeffizienten an dieser Stelle betrachtet und wiederum das Maximum gewählt. Der so ermittelte Punkt der Ridge Line (in Abbildung 3.11 mit einem Stern versehen), d.h. seine Position und Skale, wird als optimale Beschreibung des chromatographischen

[6] R-Package zur Centroidisierung von FTICR-Spektren.

Peaks betrachtet. Der Sonderfall von chromatographisch überlappenden Peaks wird in einem eigenen Abschnitt 3.3.3 behandelt.
Die Koeffizienten dieser Skale können nunmehr benutzt werden, um die Peakgrenzen festzustellen. Ausgehend von der Position des lokalen Maximums wird dazu ein beidseitiger Abstieg auf den Koeffizienten bis zum nächsten lokalen Minimum durchgeführt. Der Abstieg auf den Waveletkoeffizienten der Skale (entsprechend dem gefilterten Chromatogramm mittels des äquivalenten angepassten Filters) bietet gegenüber dem Abstieg auf den Rohdaten den Vorteil, dass durch die Filterung rauschbedingte lokale Minima des Peaks entfernt wurden und die Peakgrenzen leichter bestimmt werden können.
Die Position des lokalen Maximums auf der Skale wird als als Peakmittelpunkt $\widehat{\mathrm{rt}}$ betrachtet, die durch den Abstieg gefundenen Positionen $\mathrm{rt}_{\min}$ und $\mathrm{rt}_{\max}$ als Grenzen des chromatographischen Peaks. Weiterhin werden das Maximum $\mathrm{I}_{\max}$ sowie die Summe I der Intensitätswerte im Intervall $(\mathrm{rt}_{\min}, \mathrm{rt}_{\max})$ bestimmt.
Optional kann an dieser Stelle die Anpassung einer Gauss-Funktion $g(t) = h\mathrm{e}^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2}$ mit Streckungsparameter h an den chromatographischen Peak durchgeführt werden[7]. Die bisher gesammelten Information können als Startwerte für die Anpassung benutzt werden ($\mu = \widehat{\mathrm{rt}}$, $\sigma = (\mathrm{rt}_{\max} - \mathrm{rt}_{\min})/2$, $h = \mathrm{I}_{\max}$). Die Anpassung der Gauss-Funktion wird mittels Gauss-Newton-Verfahren (implementiert in der R-Funktion *nls*) im Intervall $(\mathrm{rt}_{\min}, \mathrm{rt}_{\max})$ durchgeführt. Neben den Parametern μ, σ, h wird auch die Wurzel aus dem mittleren quadratischen Fehler (*root mean square error*, RMSE) zwischen den Intensitätswerten und der angepassten Gauss-Funktion als Maß für die Anpassungsgüte berechnet. Der Parameter μ kann nun als neue Schätzung für den Peakmittelpunkt $\widehat{\mathrm{rt}}$ verwendet werden, der RMSE-Wert dient als Maß für die Peakqualität im Sinne der Gaussförmigkeit des Peaks. Da der Anpassungsschritt für tausende Peaks relativ viel Rechenzeit erfordert und nicht zwingend notwendig ist, wird er in der Praxis häufig weggelassen. Die Anpassung liefert jedoch nützliche Informationen für z.B. Statistiken bezüglich der Peakform und -qualität. Abbildung 3.12 zeigt das Ergebnis dieses Verarbeitungsschrittes für die Beispielregion.
Das Signal/Rausch-Verhältnis $\mathrm{SNR} = \frac{\mathrm{I}_{\max} - \mathrm{BL}}{\mathrm{NL}}$ des Peaks wird wie in Schritt 1 beschrieben bestimmt. Da jetzt die maximale Intensität einzelner Peaks betrachtet wird und nicht mehr die der gesamten ROI, ist es notwendig, nochmals zu überprüfen ob $\mathrm{SNR} \geq \mathrm{SNR}_{\mathrm{Thr}}$ gilt. Falls nicht, wird der Peak verworfen.

5. Bestimmung der *m/z* Feature-Parameter

Mit $\widehat{\mathrm{rt}}, \mathrm{rt}_{\min}, \mathrm{rt}_{\max}, \mathrm{I}_{\max}, \mathrm{I}$ und SNR sind nunmehr alle wichtigen Parameter bestimmt, die sich aus dem Chromatogramm ableiten lassen. Ausgehend von dem genau definierten

[7] Es existieren verschiedenste Modelle zur Beschreibung von chromatographischen Peaks, welche auch besondere Eigenschaften wie z.B. Tailing erfassen können [Felinger 1998, Di Marco und Bombi 2001]. Diese besonderen Eigenschaften waren hier nicht von Interesse, die anzupassende Funktion kann jedoch leicht ausgetauscht werden.

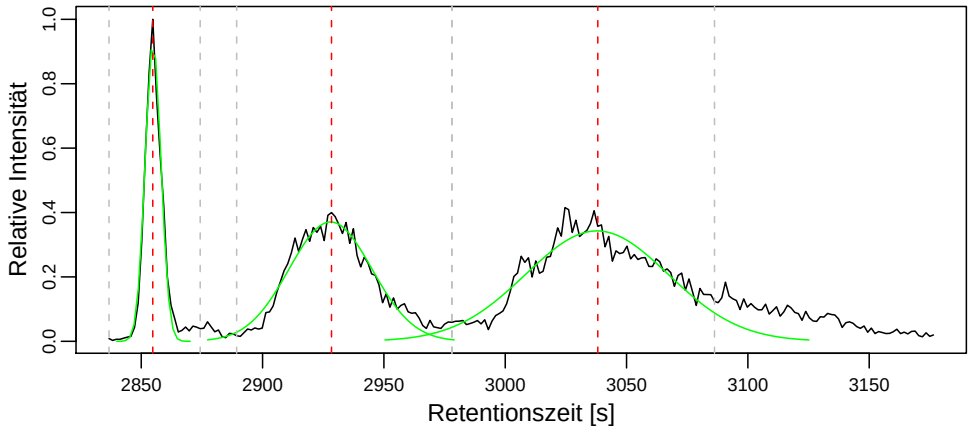

Abbildung 3.12: Peakmittelpunkte (rote Striche), Peakgrenzen (graue Striche) und angepasste Gauss-Funktionen (grün).

chromatographischen Peak können nun die Parameter in m/z-Richtung bestimmt werden, womit ab diesem Punkt die Bezeichung *Feature* verwendet werden kann.
Die Parameter m/z_{min} und m/z_{max} ergeben sich als Minimum und Maximum der m/z Werte der Region zwischen $\mathrm{rt}_{\mathrm{min}}$ und $\mathrm{rt}_{\mathrm{max}}$. Wie in Abbildung 3.13 zu sehen ist, schwanken

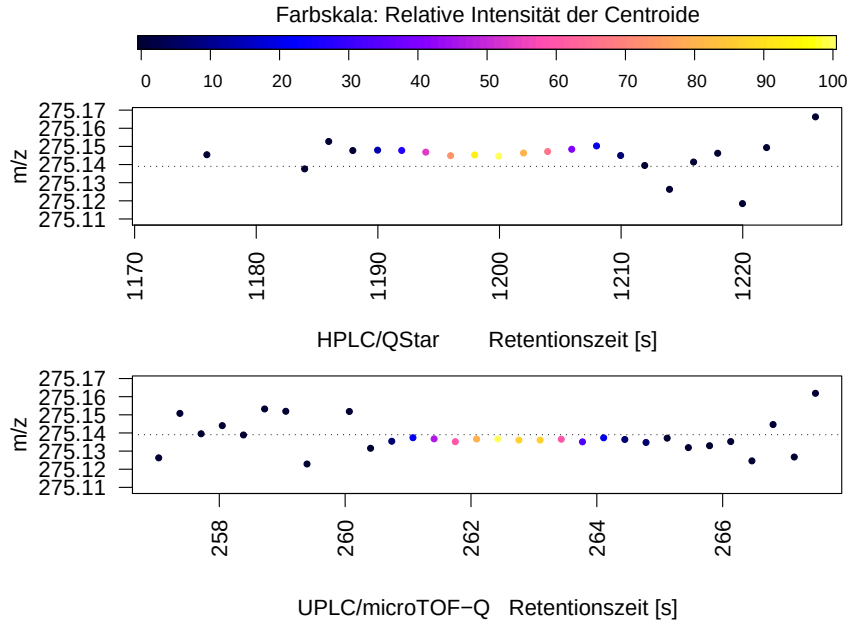

Abbildung 3.13: 10 μM IAA-Valin $[\mathrm{M+H}]^+$: Geräteabhängige Abweichung der m/z-Centroide vom Sollwert (gestrichelte Linie) bei **niedriger** Intensität. Oben: ABI/Sciex-QStar mit TDC-Detektor an einer HPLC-Säule. Unten: Bruker-microTOF-Q mit ADC-Detektor an einer UPLC-Säule.

die m/z-Centroide geräteabhängig in einem bestimmten Bereich um den m/z-Sollwert des Signals. Dabei ist die Abweichung im Allgemeinen umso geringer, je höher die Intensität des Centroids ist. Als günstig hat sich daher erwiesen, den m/z-Wert des Features $\widehat{m/z}$ über eine Mittelung der m/z-Werte der Centroide zu bilden, wobei verschiedene Varianten möglich sind:

- Einfache Mittelung der m/z-Werte innerhalb $\mathrm{rt}_{\mathrm{min}}$ und $\mathrm{rt}_{\mathrm{max}}$
- Mit den Intensitätswerten gewichtete Mittelung der m/z-Werte innerhalb $\mathrm{rt}_{\mathrm{min}}$ und $\mathrm{rt}_{\mathrm{max}}$
- Gewichtete oder ungewichtete Mittelung innerhalb eines kleineren Bereiches als $\mathrm{rt}_{\mathrm{min}}$ und $\mathrm{rt}_{\mathrm{max}}$, z.B. erst ab einer gewissen Intensitätsgrenze

Der vor allem in älteren Geräten verwendete TDC (*time-to-digital-converter*)-Detektor hat die Eigenschaft, bei hohen Intensitäten die Masse systematisch zu unterschätzen [Chernushevich et al. 2001]. Abbildung 3.14 illustriert diesen Effekt. Es existieren Ansätze zur Korrektur dieser Abweichung [Mihaleva et al. 2008]. Bei ADC (*analog-to-digital-converter*)-Detektoren tritt dieser Effekt nicht auf, dagegen kann aber gelegentlich bei sehr hohen Intensitäten eine Überschätzung der Masse beobachtet werden. Die

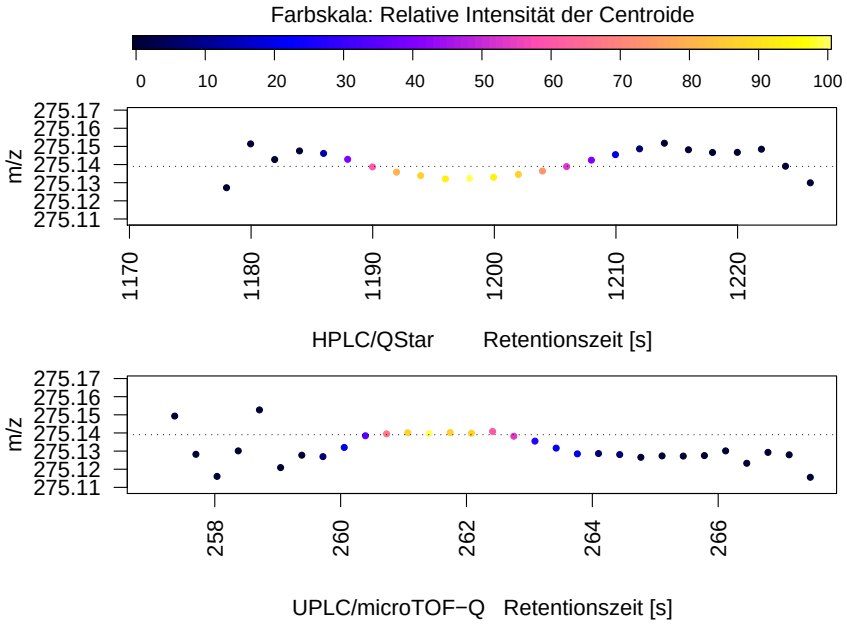

Abbildung 3.14: 100 μM IAA-Valin $[\mathrm{M+H}]^+$: Geräteabhängige Abweichung der m/z-Centroide vom Sollwert (gestrichelte Linie) bei **hoher** Intensität. Oben: ABI/Sciex-QStar mit TDC-Detektor an einer HPLC-Säule. Unten: Bruker-microTOF-Q mit ADC-Detektor an einer UPLC-Säule.

Wahl des Modells zur Berechnung des *m/z*-Wertes des Features hängt demzufolge vom Gerätetyp ab, der intensitätsgewichtete Mittelwert hat sich aber als für die meisten Geräte günstige Variante bewährt.
Um das Feature qualitativ in *m/z*-Richtung zu bewerten, wird die *m/z*-Schwankung der Centroide μ^* innerhalb von $\mathrm{rt}_{\mathrm{min}}$ und $\mathrm{rt}_{\mathrm{max}}$ in ppm berechnet. Ein gleitendes Fenster der Breite $\mathrm{p}_{\mathrm{min}}$ wird benutzt, um die $\mathrm{p}_{\mathrm{min}}$ aufeinanderfolgenden Centroide mit minimaler *m/z*-Abweichung untereinander zu finden. Kleine Werte für μ^* deuten darauf hin, dass die betreffenenden Peaks im Massenspektrum sehr gut aufgelöst wurden und demzufolge der berechnete *m/z*-Mittelwert recht genau ist. Hohe Werten für μ^* deuten auf schlecht aufgelöste Peaks hin. Hier liefert unter Umständen der *m/z*-Wert, welcher nur aus dem Spektrum mit maximaler Intensität entnommen wird, genauere Werte. Der Wert μ^* gibt auch an, welche Parametereinstellung μ des mzROI-Algorithmus ausreichen würde, um das Feature zu detektieren.
Als Ergebnis des Algorithmus liegt eine Liste von Features vor, mit Angabe des Mittelpunktes und der Grenzen in *m/z* und RT für jedes Feature, dessen Intensität sowie weiteren qualitativen Parametern.
Für die Beispielmessung wurden in den nach dem ersten Filterschritt verbliebenen 3324 ROI in 873 ROI kein Feature mit $\mathrm{SNR} \geq \mathrm{SNR}_{\mathrm{Thr}}$ detektiert, in 2406 ROI genau ein Feature, in 40 ROI zwei Features, in 4 ROI drei Features und in einer ROI fünf Features.

Betrachtung der Laufzeit
Die Laufzeit für die zweiten Phase des centWave-Algorithmus wird anhand des auf Seite 39 dargestellten Pseudocodes diskutiert. Jede der insgesamt N in der ersten Phase des Algorithmus detektierten ROI wird in der zweiten Phase auf chromatographische Peaks untersucht. Dafür wird zunächst die ROI bei einer Laufzeit von $\mathcal{O}(n)$ um eine konstante Zahl von Scans erweitert (Zeile 6), wobei die Anzahl der Scans in der erweiterten ROI mit n bezeichnet sei. Im nächsten Schritt wird eine erste Schätzung der Basislinie über den 5%-getrimmten Mittelwert der Intensitätswerte vorgenommen. Die hierfür benötigten Quantile werden über eine Sortierung (Quicksort) ermittelt. Die Laufzeit dieses Schritts beträgt damit $\mathcal{O}(n \cdot log(n))$. Die in Zeile 8 folgende Ermittlung der Intensitätswerte, die über der Basislinie liegen, wird über einen einfachen Durchlauf über alle Intensitätswerte durchgeführt. Der darauf folgende Schritt der Bestimmung von Rauschpegel und Basislinie berechnet den Mittelwert sowie die Standardabweichung. Für beide Schritte gilt eine Laufzeit von jeweils $\mathcal{O}(n)$.
Die Berechnung der CWT-Koeffizienten (Zeile 11) wird über eine schnelle Faltung durchgeführt, welche eine Laufzeit von $\mathcal{O}(n \cdot log(n))$ besitzt. Da die Faltung für jede der insgesamt s Skalen berechnet wird, ergibt sich für diesen Schritt eine Laufzeit von $\mathcal{O}(s \cdot n \cdot log(n))$. Die Ermittlung der lokalen Maxima der erhaltenen $s \cdot n$ Koeffizientenwerte wird über ein gleitendes Fenster auf jeder Skale durchgeführt. Die optimierte Variante

zur Berechnung der lokalen Maxima mittels eines gleitenden Fensters der Breite w hat eine mittlere Laufzeit von $\mathcal{O}(n)$, im ungünstigen Fall jedoch $\mathcal{O}(w \cdot n)$ [Tuszynski 2008]. Die Breite des Fensters w ist kleiner gleich n, der Aufwand für diesen Schritt somit im ungünstigen Fall $\mathcal{O}(s \cdot n^2)$. Für die Bestimmung der Ridge-Lines (Zeile 13) müssen alle gefundenen lokalen Maxima einmal betrachtet werden. Da nicht mehr als $s \cdot n$ lokale Maxima in der Koeffizientenmatrix enthalten sein können, lässt sich dieser Schritt mit $\mathcal{O}(s \cdot n)$ angeben.
Die nächsten Schritte werden für jede der insgesamt K gefundenen Ridge-Listen durchgeführt. In Zeile 16 wird die den chromatographischen Peak optimal beschreibende Skale ermittelt. Dazu wird jeder Punkt der Ridge-Line zusammen mit einem konstanten Bereich der Intensitätswerte betrachtet. Eine Ridge-Line kann als maximale Länge die Anzahl der Skalen s haben, womit sich für diesen Schritt eine Laufzeit von $\mathcal{O}(s)$ ergibt. Die beiden folgenden Verarbeitungsschritte, der Abstieg auf den Koeffizientenwerten der als optimal ermittelten Skale, sowie die Ermittlung der maximalen und summierten Intensität, besitzen jeweils eine Laufzeit von $\mathcal{O}(n)$. Die Ermittlung der Grenzen und Mittelwertbildung in m/z (Zeile 26) erfolgt ebenfalls in $\mathcal{O}(n)$. Die optionale Anpassung einer Gausskurve wird für die reguläre Feature-Detektion nicht benötigt und deren Laufzeit daher hier nicht betrachtet.
Die Laufzeit der zweiten Phase liegt damit im ungünstigen Fall in

$$\mathcal{O}\left(N \cdot s \cdot n^2 + K \cdot (s + n)\right).$$

Wie an den Werten für die Beispielmessung gezeigt, enthält eine ROI in den allermeisten Fällen nur einen chromatographischen Peak, so dass im Normalfall $K = 1$ gilt. Geht man außerdem von der mittleren Laufzeit $\mathcal{O}(n)$ für die Berechnung der lokalen Maxima aus, so ergibt sich für die zweite Phase eine mittlere Laufzeit von $\mathcal{O}\left(N \cdot s \cdot n \cdot log(n)\right)$.
Die mittlere Laufzeit für den gesamten centWave Algorithmus ergibt sich damit als

$$\mathcal{O}(C \cdot log(C) + N \cdot s \cdot n \cdot log(n)),$$

wobei C die Anzahl der Centroide in der Messung ist, N die Anzahl der in der ersten Phase detektierten ROI, n deren mittlere Länge und s die Anzahl der verwendeten Skalen. Für die Mischungen aus *A.thaliana* Samen- und Blattextrakten aus dem Datensatz LSMIX (siehe Anhang) wurden auf einer 2.2 GHz CPU mit dem Parametersatz A (siehe Abschnitt 3.4.1) Laufzeiten von rund 1 Minute pro LC/MS Messung festgestellt, wovon ca. ein Drittel auf die erste Phase (ROI-Detektion) und zwei Drittel der Laufzeit auf die zweite Phase des Algorithmus entfallen.

3.3.3 Trennung von chromatographisch überlappenden Peaks

Je nach Art der chromatographischen Trennung sowie Komplexität der LC/MS Daten können für einige Massensignale sehr eng benachbarte, oder auch überlappende chroma-

tographische Peaks beobachtet werden. Die UPLC-Technik bietet dabei im Vergleich zur HPLC eine deutliche bessere chromatographische Auflösung [Swartz 2005, Nordström et al. 2006]. Zwei chromatographische Peaks gelten als Basislinien-getrennt, wenn die gemessene Intensität zwischen den Peaks wieder auf das Niveau der Basislinie fällt. In Abbildung 3.15 sind Beispiele für nicht Basislinien-getrennte Peaks bei HPLC/MS Daten dargestellt.

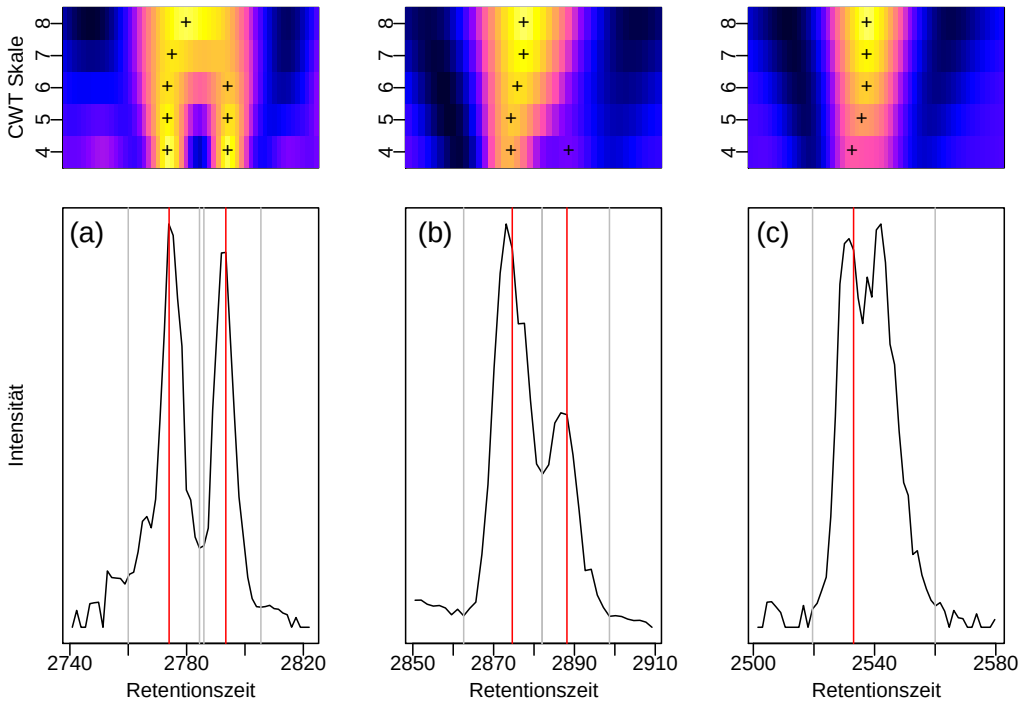

Abbildung 3.15: Ergebnisse bei chromatographisch überlappenden Peaks. Dargestellt sind drei Massensignale (v.l.n.r. m/z (796.51, 975.56, 861.52) $\pm$ 0.01) aus dem Datensatz E105. Rote Linien markieren die erkannten Zentren, graue Linien die Grenzen der chromatographischen Peaks. In der oberen Bildhälfte sind die Intensitäten der jeweiligen CWT-Koeffizienten auf den Skalen 4-8 farbkodiert dargestellt (Farbskala wie in vorherigen Abb., Skalen > 8 nicht dargestellt).

Die zwei im Teil (a) der Abbildung gezeigten, nur gering überlappenden Peaks werden mittels CWT problemlos separiert. Beide Peaks erzeugen lokale Koeffizientenmaxima auf den Skalen 4-6. Als optimal beschreibende Skale wird mit den bei Verarbeitungsschritt 4. auf Seite 42 erwähnten Regeln für beide Peaks die Skale 4 ausgewählt.
Die Fälle (b) und (c) sind bezüglich der Peakform nicht eindeutig zu interpretieren. In beiden Fällen könnte es sich um jeweils zwei überlappende Peaks handeln, möglich ist aber auch, dass es sich in Wahrheit nur um jeweils einen chromatographischen Peak handelt, dessen Form z.B. aufgrund von Intensitätsschwankungen durch Matrixeffekte (siehe

Seite 61) verfälscht wurde. In Fall (b) führt das relativ ausgeprägte Intensitätsminimum zwischen beiden Teilen des Peaks dazu, dass die Koeffizienten auf der Skale 4 für beide Teile (oder beide Peaks, je nach Interpretation) ein lokales Maximum bilden und somit zwei getrennte Peaks erkannt werden. Als beschreibende Skale wird wiederum beide Male 4 ausgewählt.
Im Fall (c) zeigt die Skale 4 nur ein lokales Maximum, es wird daher nur ein chromatographischer Peak erkannt. Da die Peakintensität im Bereich $\pm$ 2 Scans um die Position des Maximums auf Skale 4 höher ist als bei den Positionen der Maxima auf den größeren Skalen, wird auch hier als beschreibende Skale 4 ausgewählt.
Das Resultat des Algorithmus für derartige Problemfälle ist somit von der Wahl der unteren Skalengrenze abhängig. Wird hier wie für das in Abbildung 3.11 gezeigte Beispiel aus demselben Datensatz als untere Skalengrenze 6 gewählt, so werden die gering überlappenden Peaks im Fall (a) getrennt, (b) und (c) resultieren in einem Peak. Andererseits könnte auch die Trennung von Peaks wie im Fall (c) erzwungen werden, indem die untere Skalengrenze noch weiter herabgesetzt wird. Skalen < 4 erscheinen jedoch hier aufgrund der erwähnten Unsicherheiten bei der Interpretation in diesem Grenzbereich nicht sinnvoll. Allgemein werden Fälle wie (b) oder (c) relativ selten beobachtet. Um die in Abbildung 3.15 gezeigten Beispiele zu finden, wurden für den verwendeten HPLC/MS Datensatz E105 aus 5699 detektierten ROI nur solche Regionen betrachtet, die mehr als 20 Centroide enthalten. Die resultierenden 326 ungewöhnlich langen Regionen wurde manuell überprüft. Fälle vom Typ (b) oder (c) wurden dabei nur 16 mal gesichtet.

3.4 Evaluierung

Wie eingangs erwähnt, wird von einem Feature-Detektions-Algorithmus erwartet, auf der einen Seite sehr sensitiv zu sein, d.h. unter anderem auch Features mit geringer Intensität zu erkennen und auf der anderen Seite möglichst spezifisch zu arbeiten, was bedeutet falsch-positive Features (z.B. durch chemisches Rauschen verursacht) zu vermeiden.
Wünschenswert zur Evaluierung der Feature-Detektions Algorithmen wäre daher ein umfangreicher LC/MS-Datensatz, in dem alle zu findenden Features bestimmt sind. Leider steht ein solcher Datensatz nicht zu Verfügung. Um trotzdem eine Evaluierung durchführen zu können, wurde zunächst eine Mischung von relativ wenigen, aber bekannten Substanzen genutzt, um die gerätespezifischen Parameter der verglichenen Algorithmen zu optimieren. Für den eigentlichen Vergleich wurden zwei Experimente mit Verdünnungsreihen sowie Mischungen von Extrakten pflanzlicher Herkunft durchgeführt. Dazu wurde mithilfe der Algorithmen ein Referenzdatensatz von unbekannten, jedoch „verlässlichen“ Features erstellt und anschließend überprüft, inwieweit diese Features bei verschiedenen Konzentrationen bzw. in den Mischungen wiedergefundenen werden.
Wie bereits in Abschnitt 3.2.3 beschrieben, gibt es derzeit neben dem in diesem Kapitel

beschriebenen *centWave*-Algorithmus nur zwei andere Algorithmen, die für die Feature-Detektion von Metabolomik-Daten im Centroid-Modus geeignet und sowohl frei verfügbar als auch quelloffen sind: *matchedFilter* – der in XCMS ursprünglich implementierte Feature-Detektions-Algorithmus und der *centroidPicker*-Algorithmus von MZmine.
Zwei dem Autor vorliegende, kommerzielle Programme, MetAlign [MetAlign 2006] und MarkerView [MarkerView] konnten nicht in die Evaluation einbezogen werden. Beide Programme führen zwar eine Feature Detektion auf Metabolomik Daten aus, verarbeiten diese Daten jedoch intern weiter mit nachfolgenden Verarbeitungsschritten wie Alignment und Statistik. Ein Export der reinen Feature Listen, mit den Werten von m/z und Retentionszeit für jedes einzelne Feature, ist bei beiden Programmen nicht möglich.
Im Folgenden wird zunächst die Parameteroptimierung der Algorithmen beschrieben und anschließend die Evaluierung in zwei Experimenten.

3.4.1 Parameteroptimierung

Jeder der drei Algorithmen besitzt eine Anzahl von gerätespezifischen Parametern, die eingestellt werden müssen, um auf den verwendeten LC/MS-Daten gute Ergebnisse zu erzielen. Der centWave-Algorithmus benutzt der Parameter *peakwidth* ($= w_{min}, w_{max}$) um die Bereich der chromatographischen Peakbreite zu spezifizieren, den *ppm* Parameter um die zulässige Massenabweichung einzustellen, sowie *snthresh* zur Angabe des chromatographischen Signal/Rausch-Schwellwertes.
Der matchedFilter-Algorithmus besitzt einen vergleichbaren Parameter *snthresh*, die chromatographischen Peakbreite wird über den Parameter *fwhm* angegeben, der die Breite des Modellpeaks für den angepassten Filter bestimmt. Die Massenabweichung wird indirekt über die Breite der „mass slices“ mit dem Parameter *step* eingestellt.
Beim *centroidPicker* von MZmine wird in ähnlicher Art die Breite der bins mit *bin size* angegeben und zusätzlich die tolerierte Massenabweichung mit *m/z tolerance*. Desweiteren existieren fünf Parameter, die die chromatographische Domäne betreffen: *chromatographic threshold level*, *intensity tolerance*, *minimum peak duration*, *minimum peak height* und *noise level*.
Der Datensatz MM14, eine Mischung von 14 bekannten Substanzen (Details im Anhang), wurde zur Parameteroptimierung benutzt. Wie in Kapitel 4 ausführlich diskutiert, werden bei der Elektrospray-Ionisierung für jedes Molekül zahlreiche Ionen (verschiedene Addukte und Fragmente) erzeugt. Für die verwendeten Substanzen wurde im Zusammenarbeit mit einem Chemiker (Christoph Böttcher, IPB Halle) eine manuelle Annotation aller erklärbaren Features durchgeführt, die den aus den Substanzen gebildeten Addukten, Fragmenten und deren Isotopomeren entsprechen. Diese Annotation wurde auf Messungen der Einzelsubstanzen durchgeführt.
Dies ergab eine Liste von 296 annotierten Features, d.h. im Durchschnitt 21 Features

für jede der 14 Substanzen. Diese 296 Features wurden als eine Datenbasis von „echten“ Features betrachtet, die von den Algorithmen in der Mischung nach Möglichkeit detektiert werden sollten.

Anzumerken ist, dass nicht alle durch die Substanzen hervorgerufenen Features annotiert werden konnten. Die Zuordnung ist unter Kenntnis der Ionisierungsvorgänge nur für einen gewissen Teil der Features möglich (siehe Kapitel 4). In jedem Fall verbleibt eine beachtliche Anzahl von Features, die nicht ohne weiteres erklärt werden können. Dies sind wahrscheinlich unbekannte Addukte und Fragmente, oder auch durch Verunreinigungen der verwendeten Substanzen hervorgerufene Features.

Weiterhin tritt eine große Menge von „Hintergrund“-Features auf, welche auch dann beobachtet werden können, wenn nur eine Blindprobe gemessen wird. Ursache dieser Features können Weichmacher in Gefäßen und Schläuchen, Verunreinigungen in den Glasgefäßen [Mallet et al. 2006], Lösungsmittelreste, Verunreinigungen der Säule („Säulenbluten“) u.a. Substanzen sein, die aufgrund der hohen Empfindlichkeit des Massenspektrometers schon bei geringen Konzentrationen zum Teil deutliche Features bilden können. Die Autoren von [Keller et al. 2008] geben eine Liste von 394 Ionen im Bereich von m/z 50–1000 an, die als störender Hintergrund in ESI-Spektren auftreten können.

In [Trieglaff et al. 2008] werden ähnlich geartete Probleme – unbekannter Hintergrund bzw. fehlende Annotationen – bei der Feature Detektion auf Proteomik-Daten beschrieben. Die Autoren versuchen, dem Mangel an vollständig annotierten Daten zwecks Evaluierung der Algorithmen mit künstlich generierten Proteomik-LC/MS-Daten zu begegnen. Momentan ist allerdings kein Programm verfügbar, mit dem Metabolomik-LC/MS-Daten simuliert werden können, daher konnte keine solche Evaluierung zusätzlich zu der auf den echten Daten durchgeführt werden.

Eine manuelle Überprüfung der Features in der gemessenen Mischung von 14 Substanzen ergab, dass darin 122 der insgesamt 296 möglichen Features als deutlich erkennbar bezeichnet werden können. Die restlichen 174 Features sind in der Messung nur sehr schwach, z.T. kaum erkennbar ausgeprägt. Diese Diskrepanz ist dadurch zu erklären, dass die manuelle Annotation auf den Messungen der Einzelsubstanzen durchgeführt wurde, die in einer höheren Konzentration (100μM) als in der Mischung (20μM) vorlagen.

Die 122 deutlich erkennbaren Features sollten von den Algorithmen möglichst vollständig erkannt werden. Alle anderen (neben den 296 bekannten) von den Algorithmen gemeldeten Features wurde als falsch positive Features betrachtet. Diese Features können wie beschrieben entweder aufgrund von Verunreinigungen entstanden sein (in diesem Fall handelt es sich um „echte“, aber nicht erklärte Feature) oder aber auf einer Falschmeldung durch den Algorithmus beruhen (z.B. Interpretation von Rauschen als Feature). Eine gewisse Menge solcher nicht ohne weiteres erklärbarer Features muss daher in jedem Fall toleriert werden, jedoch sollten die Parameter so gewählt werden, dass die Menge dieser Features

nicht zu groß wird, da sonst das Ergebnis für den Anwender schwer zu interpretieren ist. Ein Vorversuch ergab, dass ca. 70-100 dieser 122 Features von den Algorithmen relativ zuverlässig detektiert werden. Wählt man die Parameter jedoch so, dass mehr Features gefunden werden, steigt die Anzahl der sonstigen (als falsch positiv bewerteten) Features sehr stark an. Etwas mehr als 100 Features werden von allen Algorithmen detektiert, wenn gleichzeitig bis zu 450 sonstige Features zugelassen werden.
Als Ergebnis des Vorversuchs wurde festgelegt, dass die Parameteroptimierung zunächst dahingehend durchgeführt wird, dass möglichst viele der 122 vorgegebenen Features detektiert werden sollen – unter der Bedingung, dass gleichzeitig die Anzahl der sonstigen Features eine Grenze von 450 nicht übersteigt. Dieser Parametersatz (A) stellt einen Kompromiss zwischen der Detektion erwarteter, „echter“ Features und einer möglichst geringen Anzahl von falsch positiven Features dar.
Weiterhin wurde ein zweiter Parametersatz (B) mit dem Ziel höherer Sensitivität erstellt, unter der Vorgabe *mindestens* die 122 vorgegebenen Features zu finden, unter der Bedingung, dass die Anzahl der sonstigen Features eine Grenze von 1000 nicht übersteigt.

	A		B	
Algorithmus	MM14 Features	Sonstige Features	MM14 Features	Sonstige Features
centWave	115	443	136	898
matchedFilter	114	425	144	917
MZmine	107	442	124	907

Tabelle 3.1: Anzahl der mit den optimierten Parametereinstellungen A und B von den Algorithmen detektierten Features im MM14 Datensatz.

Die Optimierung wurde für jeden Algorithmus über einen breiten Parameterbereich ausgeführt, wobei die Charakteristik des Datensatzes (chromatographische Peakbreiten, Genauigkeit des Massenspektrometers) und die von den jeweiligen Autoren empfohlenen Einstellungen als Ausgangspunkt dienten. Die getesteten Parameterbereiche und die gewählten Werte sind in Tabelle 8.1 im Anhang dargestellt. Tabelle 3.1 zeigt die Anzahl der mit optimierten Parametereinstellungen erhaltenen Features.
Um die Herkunft der sonstigen (als falsch positiv bewerteten) Features zu überprüfen, wurde mit den optimierten Parametern auch eine Feature-Detektion auf Blindproben durchgeführt (Tabelle 3.2). Als Blindprobe dienten Messungen des für die pflanzlichen Extrakte (siehe Abschnitt 3.4.6) verwendeten Lösungsmittels: 30-prozentiges Methanol. Dabei wurden zwischen 134 und 378 Features detektiert. Die möglichen Ursachen für diese „Hintergrund“-Features wurden bereits diskutiert.
Die höhere Anzahl der in der Markermischung (Tabelle 3.1) gefundenen falsch positi-

	centWave	matchedFilter	MZmine
Parametersatz A	169	134	172
Parametersatz B	325	299	378

Tabelle 3.2: Anzahl von Features, die in Blindproben von den Feature-Detektions Algorithmen gefunden werden. Angegeben ist der gerundete Mittelwert für zehn Messungen von Blindproben.

ven Features dürfte somit vor allem auf Verunreinigungen der verwendeten Substanzen zurückzuführen sein. Die Reinheit dieser Einzelsubstanzen wird je nach Hersteller meist mit 95% – 99% angegeben. Da die Markermischung mit relativ hoher Konzentration (20 μM) angefertigt wurde, ist es sehr wahrscheinlich, dass auch etliche der Verunreinigungen oberhalb der Detektionsgrenze liegen, d.h. mehr oder weniger deutliche Features hervorrufen.

3.4.2 Maße für die Bewertung der Algorithmen

Betrachtet man die zu detektierenden Features als in einer Messung „versteckt“, so kann die Feature-Detektion als ein Prozess des Information Retrieval (IR) gesehen werden. Zur Einschätzung der Leistung eines Feature-Detektions Algorithmus können daher die aus dem IR bekannten Maße von *Precision* und *Recall* verwendet werden.
Der Recall Wert misst den Anteil der durch eine Anfrage zurückgelieferten, relevanten Objekte und der Gesamtanzahl der relevanten Objekte:

$$\text{Recall} = \frac{|\,\{\text{Relevant}\} \cap \{\text{Zurückgeliefert}\}\,|}{|\,\{\text{Relevant}\}\,|} \quad ,$$

während die Precision den Anteil der erhaltenen und relevanten Objekte in Bezug zur Gesamtanzahl der erhaltenen Objekte misst:

$$\text{Precision} = \frac{|\,\{\text{Relevant}\} \cap \{\text{Zurückgeliefert}\}\,|}{|\,\{\text{Zurückgeliefert}\}\,|} \quad .$$

Sei die Gesamtanzahl der Features, die von einem Algorithmus in einer Messung detektiert wurde mit N bezeichnet, die Anzahl der echten Features die detektiert wurden mit TP und die Gesamtanzahl der zu findenden echten Features mit NP, so können Recall und Precision eines Feature-Detektions Vorgangs angegeben werden mit:

$$\text{Recall} = \frac{\text{TP}}{\text{NP}} \quad \text{und} \quad \text{Precision} = \frac{\text{TP}}{\text{N}} \; .$$

Ein perfekter Feature-Detektions Algorithmus würde mit beiden Maßen 100% erreichen. Falsch positive Features führen zu geringeren Precision-Werten, während falsch negative (nicht gefundene echte Features) den Recall-Wert senken.

Da nur die Interpretation beider Maße gemeinsam eine sinnvolle Aussage über die Leistungsfähigkeit des Algorithmus ermöglicht, wird für die Bewertung der Algorithmen auch der *F-score* als harmonischer Mittelwert von Precision(P) und Recall(R) verwendet [Rijsbergen 1979]

$$\text{F-score} = \frac{2 \cdot \text{R} \cdot \text{P}}{\text{R} + \text{P}}$$

der außerdem eine kompakte Repräsentation der gemessenen Werte erlaubt.

3.4.3 Beschreibung der Experimente

Die Ziel der beiden folgenden Experimente ist, zu überprüfen, wie gut eine komplexe Mischung von Substanzen zum einen in verschiedenen Verdünnungsstufen detektiert werden kann und zum anderen, wenn diese Mischung in einer zweiten, ebenfalls komplexen Mischung „versteckt“ wird. Da keine Mischung aus genügend vielen bekannten Substanzen zur Verfügung stand, wurde dafür eine Mischung von unbekannten Substanzen benutzt, indem pflanzliche Samen- und Blattextrakten zunächst getrennt in verschiedenen Verdünnungsstufen und dann als Mischung (in verschiedenen Verhältnissen) gemessen wurden.
In zwei Schritten wurden reproduzierbare und von der Mehrzahl der Algorithmen detektierte Features aus den *unverdünnten* Samen- und Blattextrakten gewonnen, welche anschließend als Referenzdatensatz benutzt wurden, um damit Recall und Precision der Algorithmen in den Verdünnungsstufen und Mischungen der beiden Extrakte zu untersuchen. Dafür werden alle im Referenzdatensatz enthaltenen Features als „echte“, zu detektierende Features angenommen (TP). Alle sonstigen, von den Algorithmen zurückgelieferten Features werden als falsch positiv beurteilt.
Da die beiden im Optimierungsschritt gefundenen Parametersätze eine sehr unterschiedliche Anzahl von Features detektieren, war es für die Auswertung erforderlich, für jeden Parametersatz einen eigenen Referenzdatensatz zu erstellen. Die mit den Parametern A erstellte Variante fällt deutlich kleiner aus, als die mit dem Parametersatz B erstellte. Dieser Referenzdatensatz kann somit auch nur zur Auswertung mit den Parametern A dienen. Bei der Untersuchung mit den Parametern B würden alle nicht darin enthaltenen Features als falsch positiv bewertet werden, obwohl sie evtl. im Referenzdatensatz B enthalten sind und damit doch „echte“ Features darstellen, lediglich bezüglich eines anderen Schwellwertes. Die erhaltene Precision wäre damit verfälscht. Andersherum, bei der Untersuchung der mit Parametersatz A detektierten Features auf dem Referenzdatensatz B, würde die Aussage des Recalls verfälscht, da dann viele Features des Referenzdatensatzes aufgrund des niedriger eingestellten Schwellwertes nicht detektiert werden. Aus diesen Gründen wurden die Auswertung für jeden Parametersatz mit dem jeweiligen Referenzdatensatz vorgenommen.

Für die Erstellung der Referenzdatensätze und die beiden Experimente wurden Mischungen bestehend aus Lösungsmittel, Samen- und Blattextrakten (von *Arabidopsis thaliana*) in den folgenden Volumenverhältnissen gemischt: 0/100/0, 25/75/0, 50/50/0, 75/25/0, 0/0/100, 25/0/75, 50/0/50, 75/0/25, 0/75/25, 0/50/50, 0/25/75, 100/0/0 (Lösungsmittel/Samen/Blatt). Die 12 Proben wurden in jeweils zehn technischen Replikaten gemessen (Datensatz LSMIX, technische Details im Anhang).

3.4.4 Erstellung der Referenzdatensätze

Die Erstellung der Referenzdatensätze wurde in zwei Schritten vorgenommen: Im ersten Schritt wurden nur reproduzierbar detektierte Features zurückbehalten und im zweiten Schritt wurden die erhaltenen Feature-Listen verglichen und nur solche Features zurückbehalten, die von mindestens zwei der drei Algorithmen detektiert wurden.

Reproduzierbar detektierte Features

Um eine verlässliche Basis von samen- und blattspezifischen Features zu bilden und eher zufällig auftretende, störende Signale auszuschließen, wurden im ersten Schritt nur solche Features ausgewählt, die in mindestens sieben der zehn technischen Replikate von den jeweiligen Algorithmen detektiert wurden. Dazu wurde die Feature-Detektion auf unverdünnten Samen- und Blattextrakten (Proben Lösungsmittel/Samen/Blatt: 0/100/0 und 0/0/100) mit den drei Algorithmen unter Benutzung der beiden optimierten Parametereinstellungen (A & B) durchgeführt. Die resultierenden Feature-Listen wurden anschließend separat unter Benutzung der XCMS Funktion *group* aligniert (Details zum Alignment-Algorithmus in Kapitel 5). Die dafür verwendeten Alignment-Parameter *mzwid* = 0.05 und *bw*=2 sind durch die Kenntnis der auftretenden Abweichungen und Erfahrungen mit dem Alignment-Algorithmus begründet und stellen für diese Gerätekombination die standardmäßig verwendete Einstellung dar. Als Repräsentant wird für jede Feature-Gruppe nach dem Alignment der Median bezüglich der *m/z* Werte als auch der Retentionszeit von den alignierten Features gewählt. Das Alignment wurde unter der erwähnten Einschränkung durchgeführt, dass jede der gebildeten Feature-Gruppen mindestens sieben Features enthält (Parameter *minfrac*=0.7).

Die Anzahl der alignierten Features sinkt deutlich, je mehr Messungen für das Vorkommen eines Features gefordert werden (siehe Abbildung 3.16). Dieser Effekte ist neben möglichen Alignment-Problemen darauf zurückzuführen, dass die Intensitäten in den Messungen leicht variieren und damit Features die in einer Messung detektiert werden, in einer anderen unterhalb des vorgegebenen Signal/Rausch-Verhältnisses liegen. Eine weitere Erklärung ist das Auftreten von Störsignalen mit zufälliger Natur, derartige Features sind folglich nicht reproduzierbar. In Standardexperimenten werden daher mindestens drei bis vier technische Replikate pro Probe verwendet.

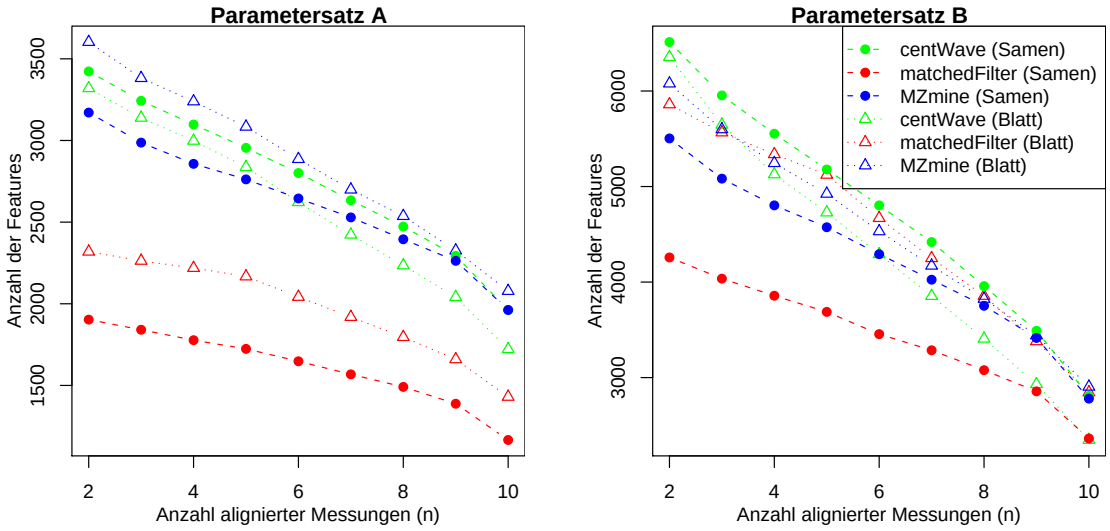

Abbildung 3.16: Anzahl der alignierten Features aus zehn technischen Replikaten, wenn das Vorkommen eines Features in mindestens n Replikaten gefordert wird.

Algorithmus	Samen-extrakt	Blatt-extrakt
centWave	2643	2423
matchedFilter	1568	1919
MZmine	2529	2699

Tabelle 3.3: Parametersatz A: Anzahl alignierter Features für n=7 technische Replikate.

Algorithmus	Samen-extrakt	Blatt-extrakt
centWave	4419	3854
matchedFilter	3286	4250
MZmine	4025	4171

Tabelle 3.4: Parametersatz B: Anzahl alignierter Features für n=7 technische Replikate.

Um die resultierenden Feature-Listen (als Ausgangspunkt für den folgenden Schritt) nicht zu stark zu dezimieren, wurden für dieses Filterkriterium sieben, jedoch nicht zehn Messungen gewählt.

Die Anzahl der auf diese Art reproduzierbaren Features ist in den Tabellen 3.3 und 3.4 gezeigt. Aufgrund der sensitiveren Einstellungen in Parametersatz B werden damit bis doppelt so viele Features wie mit Parametersatz A detektiert und auch aligniert.

Schnittmengenbildung

Im zweiten Schritt wurden die für jeden Algorithmus aus dem vorherigen Schritt erhaltenen Feature-Listen verglichen und nur solche Features zurückbehalten, die von mindestens zwei der drei Algorithmen detektiert wurden. Da die einzelnen Algorithmen verschiedene Methoden verwenden, um den Mittelpunkt eines Features anzugeben, wurde beim Vergleich eine Abweichung von 0.015 m/z und 5 Sekunden akzeptiert. Diese Werte beruhen auf einer subjektiven Einschätzung der maximalen absoluten Abweichungen bei der

Ermittlung der Featuremittelpunkte durch die Algorithmen. Bei größeren Abweichungen wurden die Features als nicht zusammengehörig betrachtet, zumal etwaige Abweichungen in einzelnen Messungen bereits durch die Benutzung des Repräsentanten der Feature-Gruppe nach dem Alignment kompensiert sein sollten. Mehrdeutigkeiten beim Vergleich wurden mit den verwendeten Toleranzwerten nicht beobachtet. Dieser Prozess der Schnittmengenbildung wurde für alle aus dem vorherigen Schritt erhaltenen Feature-Listen durchgeführt.

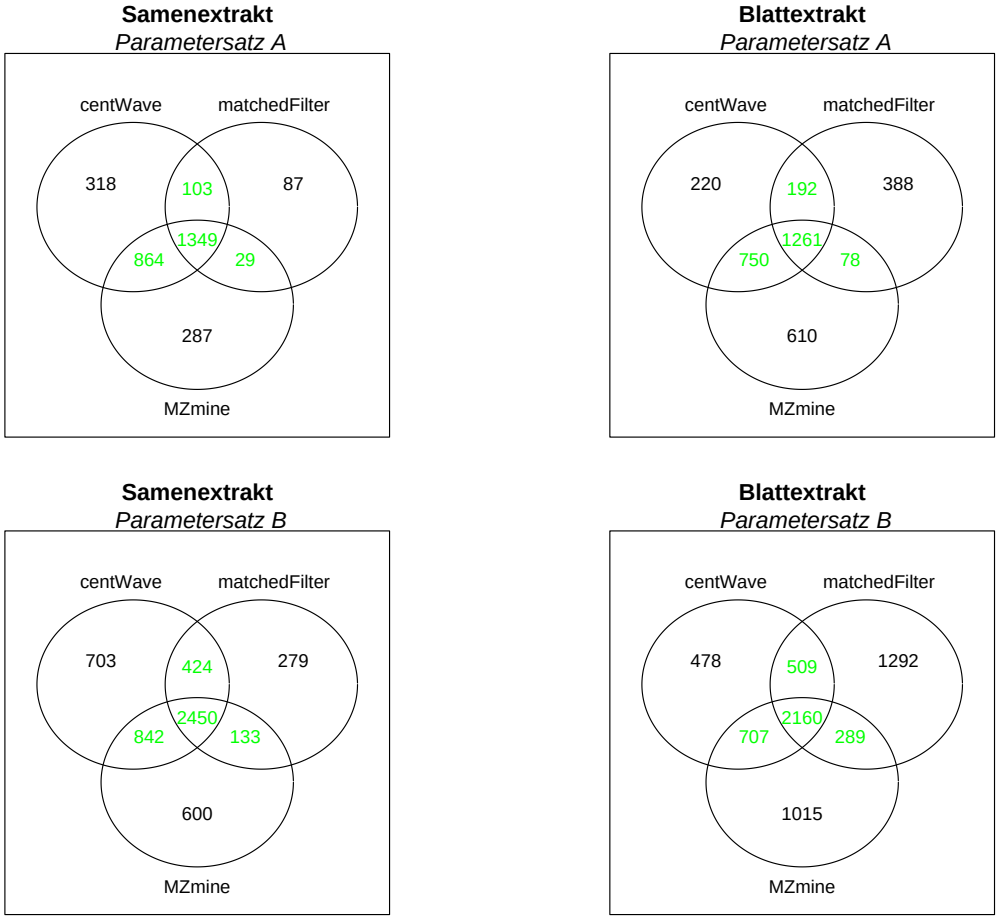

Abbildung 3.17: Venn Diagramme zur Anzahl der Features, die in Samen- und Blattextrakten von den drei Algorithmen reproduzierbar detektiert wurden. Die Features aus den gebildeten Schnittmengen (grün markiert) wurden zur Erstellung der Referenzdatensätze benutzt.

Abbildung 3.17 zeigt die Ergebnisse dieses Filterungsschritts als Venn-Diagramme. Zu beobachten ist, dass in jedem der vier Fälle die Schnittmenge zwischen centWave und MZmine den größten Anteil an der Gesamtmenge der resultierenden Features liefert. Ebenfalls in allen Fällen hat die Schnittmenge zwischen centWave und matchedFilter den nächstkleineren Anteil und demzufolge die zwischen matchedFilter und MZmine den geringsten Anteil am Ergebnis. Interpretiert man diesen Anteil gemeinsam detektierter Features als ein Maß für die Ähnlichkeit zwischen Algorithmen, so ähneln sich centWave

und MZmine bezüglich der Feature-Detektion am meisten und MZmine und matchedFilter am wenigsten.

Weiterhin lässt sich beobachten, dass ein recht hoher Anteil der insgesamt detektierten Features nur von einem Algorithmus gefunden werden (zwischen 30 und 76% bezogen auf die Anzahl der Features in den jeweiligen Schnittmengen). Für diese nicht übernommenen und damit indirekt als falsch positiv bewerteten Features ist jedoch kein eindeutiger Trend bezüglich der einzelnen Algorithmen erkennbar.

Der resultierende Referenzdatensatz für den Parametersatz A (als Referenzdatensatz A bezeichnet) enthält 2345 Features für den Samen- und 2281 Features für den Blattextrakt. Referenzdatensatz B enthält 3849 Features für den Samen- und 3665 Features für den Blattextrakt.

3.4.5 Experiment 1 : Evaluierung auf Verdünnungsreihen

Das erste Experiment beinhaltet die Evaluierung der Algorithmen auf drei Verdünnungsstufen von Samen- und Blattextrakten mittels der auf den unverdünnten Extrakten erstellten Referenzdatensätze. Die Herausforderung für die Algorithmen besteht bei diesem Experimentdesign darin, die bei hoher Konzentration deutlich ausgeprägten und damit einfacher zu detektierenden Features auch in den Verdünnungsstufen wiederzufinden.

Verwendet wurden die Messungen 25/75/0, 50/50/0, 75/25/0 und 25/0/75, 50/0/50, 75/0/25 (Lösungsmittel/Samen/Blatt) aus dem Datensatz LSMIX. Für jeden Parametersatz wurden jeweils Recall und Precision gemessen und daraus der F-score berechnet. In den graphischen Darstellungen werden jeweils F-score und Recall gezeigt. Alle gemessenen Werte, inklusive der Precision, sind auch in tabellarischer Form im Anhang (Seite 118) wiedergegeben.

Parametersatz A

Unter Verwendung des Parametersatzes A wurde die Feature-Detektion mit den drei Algorithmen auf jeweils 10 technischen Replikaten der Messungen durchgeführt. Abbildung 3.18 zeigt die ermittelten F-score und Recall-Werte bezüglich der im Referenzdatensatz A erfassten Features. Die zugehörigen Precision-Werte sind im Anhang (Tabelle 8.2, Seite 118) angegeben.

Abhängig von der Verdünnungsstufe wurden zwischen 41 und 90 % der Features aus dem Referenzdatensatz detektiert. Der Recall-Wert von centWave ist in diesem Experiment für fünf von sechs Verdünnungsstufen höher als der der anderen beiden Algorithmen, der F-score ist in jedem Fall besser. Der matchedFilter-Algorithmus liefert hier insgesamt die schlechtesten Ergebnisse.

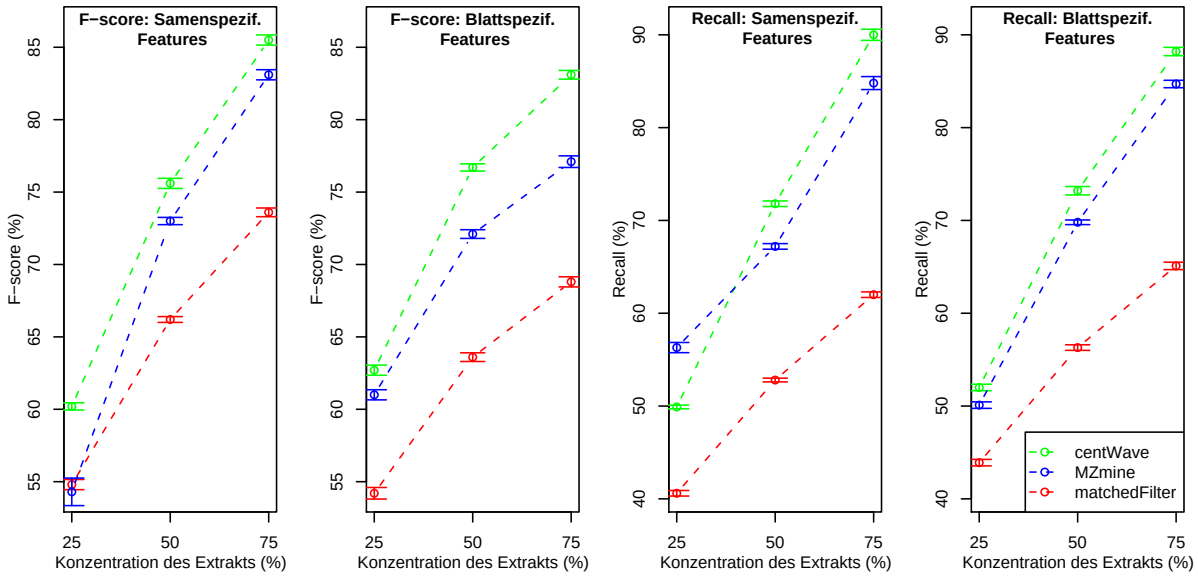

Abbildung 3.18: F-score und Recall für drei Verdünnungsstufen von Samen- und Blattextrakten bezüglich der im Referenzdatensatz A erfassten Features. Die Fehlerbalken zeigen die Standardabweichung der ermittelten Werte für die zehn technischen Replikate jeder Probe.

Parametersatz B

Unter Verwendung des Parametersatzes B wurden in den Verdünnungsstufen zwischen 46 und 88 % der Features aus dem deutlich größeren Referenzdatensatz B von den Algorithmen detektiert. Die ermittelten F-score und Recall-Werte sind in Abbildung 3.19 dargestellt. Die Precision-Werte sind im Anhang (Tabelle 8.5, Seite 119) angegeben.

Der matchedFilter-Algorithmus zeigt mit diesem Parametersatz bessere Ergebnisse als mit Parametersatz A. Die durchschnittlichen F-score und Recall-Werte des centWave Algorithmus sind in allen Fallen höher als die der anderen Algorithmen.

3.4.6 Experiment 2 : Evaluierung auf komplexen Mischungen

Das zweite Experiment umfasst die Evaluierung der Algorithmen auf Mischungen aus Samen- und Blattextrakten. Die Extrakte wurden in drei unterschiedlichen Verhältnissen kombiniert: 0/75/25, 0/50/50, 0/25/75 (Lösungsmittel/Samen/Blatt, Teil des Datensatzes LSMIX). Die Herausforderung für die Algorithmen besteht auch hier darin, die Features aus dem Referenzdatensatz bei geringeren Konzentrationen zu detektieren, jedoch sind diese jetzt zusätzlich in einem komplexen Hintergrund „versteckt“. Dabei sind zwei verschiedene Sichtweisen möglich. In der ersten wird der Blattextrakt als „Hintergrund“ betrachtet und der Recall bezüglich des darin „versteckten“ Samenextrakts untersucht, in der zweiten genau umgekehrt.

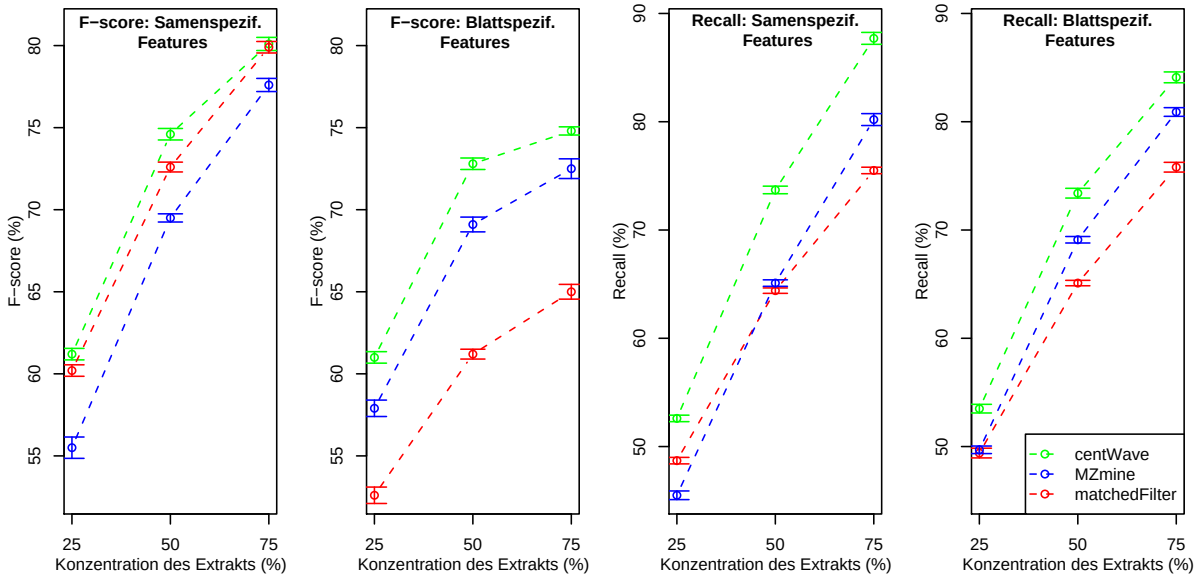

Abbildung 3.19: F-score und Recall für drei Verdünnungsstufen von Samen- und Blattextrakten bezüglich der im Referenzdatensatz B erfassten Features. Die Fehlerbalken zeigen die Standardabweichung der ermittelten Werte für die zehn technischen Replikate jeder Probe.

Beide Sichtweisen wurden untersucht und zusätzlich auch Recall, Precision und F-score in Bezug auf die Vereinigungsmenge der samen- und blattspezifischen Features betrachtet. Wie in Experiment 1 wurden auch hier beide Parametersätze in Verbindung mit den jeweiligen Referenzdatensätzen benutzt und Messungen auf jeweils 10 technischen Replikaten jeder Mischung durchgeführt. Für die graphischen Darstellungen werden wiederum F-score und Recall verwendet. Alle gemessenen Werte, inklusive der Precision, sind nochmals in tabellarischer Form im Anhang (Seite 121) gezeigt.

Parametersatz A

Die Feature-Detektion wurde mit den drei Algorithmen unter Verwendung des Parametersatzes A durchgeführt. Untersucht wurden zunächst Recall und Precision in Bezug auf die Vereinigungsmenge der im Referenzdatensatz A erfassten Features. Abbildung 3.20 (linke Hälfte) zeigt die ermittelten F-score und Recall-Werte. Detektiert wurden zwischen 42 und 63 % der Features aus der betrachteten Vereinigungsmenge von samen- und blattspezifischen Features. Sowohl Recall als auch F-score des centWave-Algorithmus sind dabei deutlich höher als die von MZmine und matchedFilter erzielten Werte.
Zur Betrachtung von allein den samen- oder blattspezifischen Features mit dem Blatt- bzw. Samenextrakt als „Hintergrund“ wurden nur die jeweils spezifischen Features aus dem Referenzdatensatz A bewertet und die Recall-Werte berechnet (Abbildung 3.20, rechte Hälfte). Zu beachten ist, dass 550 Features aus dem Referenzdatensatz A so-

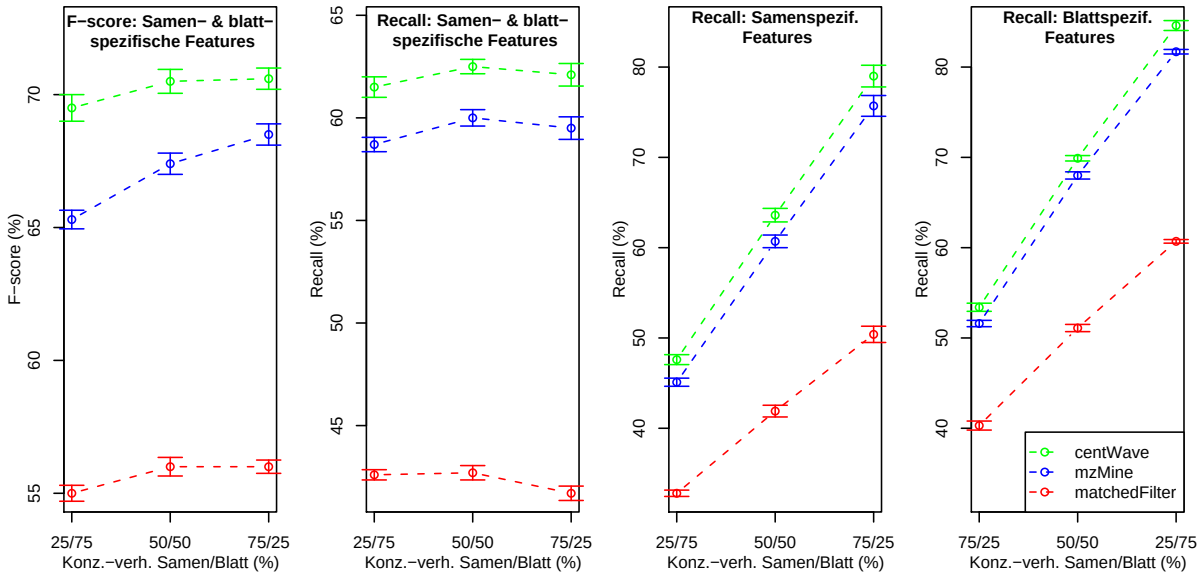

Abbildung 3.20: F-score und Recall bezüglich der Vereinigungsmenge der im Referenzdatensatz A erfassten samen- und blattspezifischen Features (linke Hälfte der Abbildung), sowie Recall-Werte für nur die samen- bzw. blattspezifischen Features (rechte Hälfte der Abbildung). Gemessen wurden Mischungen von Samen- und Blattextrakten in unterschiedlichen Konzentrationsverhältnissen. Die Fehlerbalken zeigen die Standardabweichung der ermittelten Werte für die zehn technischen Replikate jeder Probe.

wohl im Samen- als auch im Blattextrakt vorkommen, und daher in beiden Sichtweisen in die Berechnung der Recall-Werte eingehen. Die Recall-Werte liegen dabei insgesamt um durchschnittlich rund 3 (Blatt) bis 9 (Samen) Prozent niedriger als bei den ungemischten Extrakten in Experiment 1. Abgesehen von auftretenden Problemen bei der Feature-Detektion aufgrund der höheren Featuredichte ist dies auch durch auftretende *Matrixeffekte* (Veränderungen der Ionisierungseffizienz von Substanzen in Gegenwart koeluierender Substanzen) [Böttcher et al. 2007] zu erklären. Für diese Auswertung der samen- oder blattspezifischen Features kann sinnvollerweise nur der Recall-Wert gemessen werden, da die als falsch positiv bewerteten, sonstigen Features in diesem Fall auch die jeweils andere Teilmenge des Referenzdatensatzes enthalten.

Parametersatz B

Die Ergebnisse unter Verwendung des Parameter- und Referenzdatensatz B sind ähnlich. Aus der Vereinigungsmenge von samen- und blattspezifischen Features des Referenzdatensatzes wurden zwischen 49 und 62 % durch die drei Algorithmen detektiert. Die erzielten F-score und Recall-Werte von centWave sind auch hier deutlich höher als die von MZmine und matchedFilter (Abbildung 3.21, linke Hälfte).

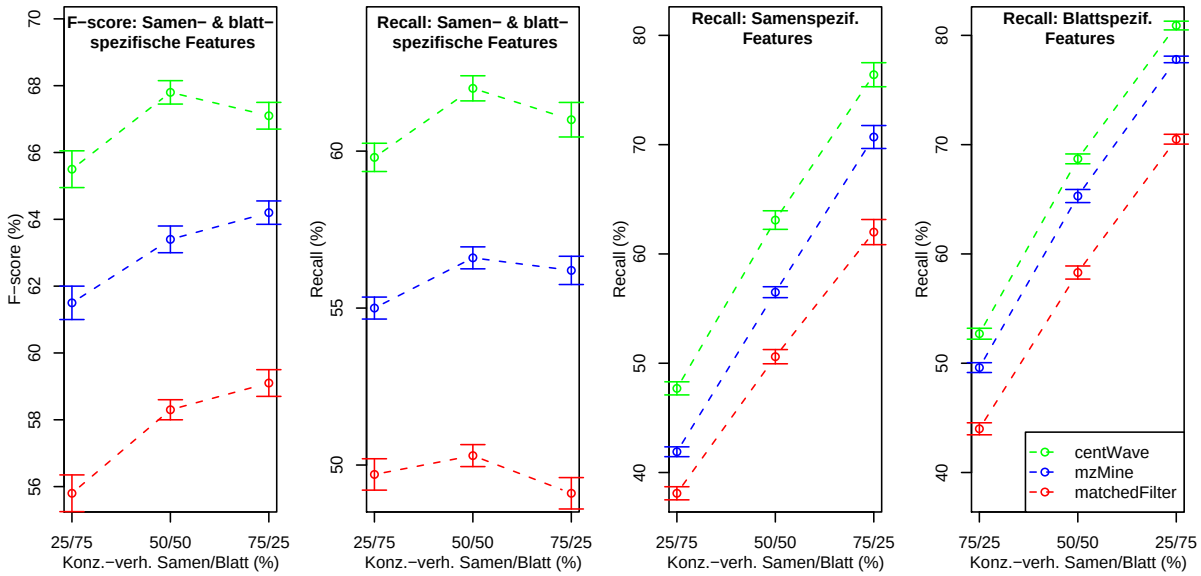

Abbildung 3.21: F-score und Recall bezüglich der Vereinigungsmenge der im Referenzdatensatz B erfassten samen- und blattspezifischen Features (linke Hälfte der Abbildung), sowie Recall-Werte für nur die samen- bzw. blattspezifischen Features (rechte Hälfte der Abbildung). Gemessen wurden Mischungen von Samen- und Blattextrakten in unterschiedlichen Konzentrationsverhältnissen. Die Fehlerbalken zeigen die Standardabweichung der ermittelten Werte für die zehn technischen Replikate jeder Probe.

Es wurden wiederum auch die Recall-Werte, bezogen auf jeweils samen- oder blattspezifische Features aus dem Referenzdatensatz B berechnet (Abbildung 3.21, rechte Hälfte). 865 Features kommen dabei in beiden Featurelisten vor. Der Recall lag um durchschnittlich rund 4 (Blatt) bis 10 (Samen) Prozent niedriger als bei den ungemischten Extrakten in Experiment 1. Die Recall-Werte von centWave sind abermals höher als die von MZmine und matchedFilter erreichten.

Laufzeit

Die Laufzeiten der einzelnen Algorithmen sind in Tabelle 3.5 angegeben. Mit dem Parametersatz A benötigt centWave nur rund eine Minute für die Feature-Detektion einer LC/MS-Messung. Der Grund dafür ist die *prefilter*-Option, die bei diesem Parametersatz mit (2,400) höher als bei Parametersatz B gewählt ist und dadurch mehr ROI's in der ersten Phase ausfiltert, die dadurch nicht mehr in der zeitaufwendigeren zweiten Phase des Algorithmus untersucht werden. Mit dem Parametersatz B (*prefilter*=(2,200)) verringert sich dieser Vorteil und die Laufzeit von centWave liegt dann zwischen der von MZmine und matchedFilter.

Die Laufzeiten von MZmine und matchedFilter sind mit dem Parametersatz B sogar

	centWave	matchedFilter	MZmine
Parametersatz A	1.02	1.85	1.54
Parametersatz B	1.54	1.71	1.42

Tabelle 3.5: Durchschnittliche Laufzeit für die Feature-Detektion einer LC/MS-Messung in Minuten. Gemittelt wurde über zehn technische Replikate einer Blatt-/Samenextrakt Mischung (50/50) aus dem Datensatz LSMIX. Die Messungen erfolgten auf einem AMD Athlon 64 X2 4200+ (2.2GHz) mit 4GB RAM unter Linux (Ubuntu 6.06).

etwas günstiger als mit Parametersatz A. Für den matchedFilter-Algorithmus liegt die Ursache in den als Ergebnis der Optimierung u.a. etwas breiter gewählten „mass slices“ (Parameter *step* 0.025 statt 0.02), was die *m/z*-Toleranz erhöht und als Nebeneffekt die Laufzeit dadurch verringert, dass im Vergleich zu A breitere und insgesamt weniger „mass slices“ gebildet werden.
Für MZmine wurde im Verlauf der Optimierung für Parametersatz B eine geringere *minimum peak height* (300 statt 500) gewählt, jedoch gleichzeitig der *chromatographic threshold level* etwas heraufgesetzt (0.85 statt 0.8). Letzterer beeinflusst die Laufzeit des Algorithmus sehr stark, da dieser Parameter als Schwellwert für die Betrachtung der *m/z*-bins dient (siehe Abschnitt 3.2.3). Bei kleineren Werten für diesen Parameter (z.B. 0.5) wurden bei der Optimierung Laufzeiten von bis zu 10 min pro LC/MS-Messung beobachtet. Die Laufzeitverbesserung gegenüber Parametersatz A lässt sich also hier durch die Erhöhung dieses Parameters erklären.
Sowohl XCMS (enthält die Algorithmen centWave und matchedFilter) als auch MZmine verfügen über Techniken, um die Feature-Detektion vieler Messungen für die parallele Verarbeitung auf mehrere CPUs und auch mehrere Rechnern zu verteilen. MZmine nutzt Java RMI (Remote Method Invocation), während XCMS dafür auf MPI (Message Passing Interface) über Rmpi [Yu 2002] zurückgreift. Diese Optionen wurde bei der Laufzeitmessung jedoch nicht benutzt.

3.5 Zusammenfassung

Der centWave-Algorithmus benutzt einen dichteorientierten ROI-Algorithmus anstelle der verbreiteten Binning-Technik zur Erkennung von Massensignalen, in Verbindung mit einem Wavelet-basierten Ansatz zur Detektion von chromatographischen Peaks.
Zum Vergleich mit zwei anderen Feature-Detektions Algorithmen (matchedFilter und MZmine) wurde eine Evaluierung auf Verdünnungsreihen und Mischungen von *A.thaliana* Samen- und Blattextrakten durchgeführt, unter Ermittlung von Precision, Recall und F-score in Bezug auf einen vorher erstellten Referenzdatensatz.
In 29 von 30 durchgeführten Messungen des Recall-Wertes lag der von centWave erzielte

Wert höher als der von matchedFilter und MZmine. Im Mittel waren die Recall-Werte von centWave 14.3% höher als die von matchedFilter und 3.8% höher als die von MZmine. Für 17 von 18 der über Precision und Recall ermittelten F-score-Werte war der vom centWave-Algorithmus erreichte F-score höher als der der anderen beiden Algorithmen, in einem Fall gleich hoch. Der von centWave erzielte F-score war dabei durchschnittlich 9.2% und 3.7% höher als der von matchedFilter bzw. MZmine erreichte Wert.
Der centWave-Algorithmus liefert damit eine deutlich sensitivere Feature-Detektion als die anderen beiden Algorithmen, ohne dass gleichzeitig vermehrt falsch positive Features gemeldet werden. Die Laufzeit des centWave-Algorithmus ist –je nach verwendetem Parametersatz– dabei zum Teil kürzer als die von matchedFilter und MZmine, oder liegt im selben zeitlichen Bereich.

4 Annotation zusammengehöriger Features

Im positiven Ionisierungsmodus werden bei der Elektrospray Ionisierung (siehe Abschnitt 2.2) vor allem sogenannte *Quasi-Molekülionen* gebildet, wie z.B. $[M+H]^+$ und $[M+Na]^+$ mit einfacher Ladung, oder Ionen mit mehrfachem Ladungszustand wie z.B. $[M+2H]^{2+}$ oder $[M+2Na]^{2+}$. Außerdem entstehen sogenannte *Clusterionen* wie z.B. $[2M+H]^+$ oder $[2M+Na]^+$. Bei den in dieser Arbeit betrachteten *A. thaliana* Extrakten werden Molekülionen $[M]^+$ eher selten beobachtet, diese treten nur bei bestimmten Stoffklassen, wie Anthocyanen oder Cholinestern, mit intrinsischem Ladungszustand[8] auf.
Im negativen Ionisierungsmodus entstehen beispielsweise das deprotonierte Molekülion $[M-H]^-$ sowie Adduktbildungen mit Anionen wie z.B. $[M+Cl]^-$.
Weiterhin werden durch kollisionsinduzierte Dissoziation (in-source collision-induced dissociation [Bure und Lange 2003]) auch *Fragmentionen* wie beispielsweise $[M-C_6H_9O_5]^+$ (bei Glycosiden) oder $[M-C_3H_9N]^+$ (bei Cholinestern) gebildet. Zusätzlich kann die Anlagerung von Molekülen aus dem Lösungsmittel, wie z.B. CH_3OH, beobachtet werden.
Die für ein Molekül M im Massenspektrometer beobachteten Massensignale s (gemessen in m/z) ergeben sich demzufolge aus der Molekülmasse $\text{mass}(M)$ und der chemischen Modifikation durch den Ionisierungsprozeß:

$$s = \frac{n\ \text{mass}(M) + \text{mass}(\{I\}) + \text{mass}(\{L\}) - \text{mass}(\{F\})}{z} \tag{4.1}$$

wobei n $(n = 1, 2, \ldots)$ die Anzahl der Moleküle im Ion ist, $\text{mass}(\{I\})$ die Gesamtmasse der angelagerten Ionen I, $\text{mass}(\{L\})$ die Gesamtmasse von aus aus dem Lösungsmittel stammenden Molekülen L und $\text{mass}(\{F\})$ die Gesamtmasse der abgespaltenen Fragmente F. Der Ladungszustand z des gebildeten Ions ergibt sich als Summe der Ladungen aller am Bildungsprozeß beteiligten Ionen. Für die am häufigsten beobachteten Fälle der Ionenbildung wie $[M+H]^+$ oder $[M+Na]^+$ entspricht der Ladungszustand des gebildeten Ions einfach dem Ladungszustand der angelagerten Ionen, wie H^+ oder Na^+. Besitzt das Molekül M eine intrinsische Ladung, so ist diese bei der Berechnung der Gesamtladung (n-mal) zu berücksichtigen. Schließlich können auch die aus dem Lösungsmittel stammenden Moleküle schon einen Ladungszustand besitzen, der dann ebenfalls addiert werden muss.
Die beobachteten Intensitäten der einzelnen Peaks bzw. Features sind innerhalb eines bestimmten Intensitätsbereichs (*dynamic range*, siehe [Chernushevich et al. 2001]) proportional zur Bildungshäufigkeit der entsprechenden Ionen. Diese wiederum hängt von den chemischen Eigenschaften des Moleküls ab. Bei konstanten Ionisierungsbedingungen sollten die beobachteten relativen Intensitäten der einzelnen Ionen ebenfalls konstant sein,

[8] Wenn Moleküle einen Ladungszustand besitzen, ohne dass zusätzlich ein Ion angelagert wurde, so wird dies als *intrinsische Ladung* bezeichnet.

d.h. die aus einem Molekül gebildeten Ionen treten innerhalb einer Messung immer im gleichen Verhältnis auf. Diese Eigenschaft wird für die Korrelationsanalyse von Features (in Abschnitt 4.3 beschrieben) genutzt. Tabelle 4.1 zeigt Beispiele für gebildete Ionen von Biochanin A nach ESI.

	Ion	m/z	Relative Intensität (%)
1	$[M - CH_3O]^+$	253.05	1
2	$[M - CH_2]^+$	270.05	1
3	$[M+H]^+$	285.08	100
4	$[M+Na]^+$	307.06	17
5	$[3M+K+H]^{2+}$	446.09	1
6	$[2M+H]^+$	569.14	1
7	$[2M+Na]^+$	591.13	2
8	$[2M+K]^+$	607.10	3
9	$[3M+Na]^+$	875.19	1
10	$[3M+K]^+$	891.17	2

Tabelle 4.1: Beispiele für beobachtete Addukte und Fragmente für Biochanin A ($M=C_{16}H_{12}O_5$) am Bruker microTOF-Q. Die am Massenspektrometer beobachteten, auf eine maximale Intensität von 100 skalierten Intensitäten wurden auf ganzzahlige Werte gerundet.

4.1 Problemstellung

Da aufgrund des Ionisierungsprozesses jede Verbindung durch eine Vielzahl von Peaks im Massenspektrum bzw. einer Vielzahl von LC/MS-Features repräsentiert wird, ist die Interpretation eines Massenspektrums bzw. einer Featureliste nicht trivial.

Es kann nicht davon ausgegangen werden, dass der höchste Peak in einem ESI-Spektrum dem Quasi-Molekülion $[M+H]^+$ entspricht, auch wenn dies ein häufig beobachteter Fall ist. Verschiedene Möglichkeiten müssen in Betracht gezogen werden, um eine richtige Zuordnung zu treffen.

Weiterhin kann nicht angenommen werden, dass alle Peaks in einem Massenspektrum bzw. alle Features im einem Zeitfenster von einer einzigen Substanz stammen. Auch bei bestmöglicher chromatographischer Trennung kommt es bei komplexen Proben vor, dass Substanzen eine ähnliche Affinität zur stationären Phase in der Säule haben und daher zum selben Zeitpunkt eluieren. In diesen Fällen überlagern sich die Massenspektren der Substanzen. Durch Verunreinigungen aller Art (Lösungsmittel, Weichmacher aus Plastikgefäßen und -schläuchen etc.) sind oft auch weitere Signale vorhanden, die nicht eindeutig zugeordnet werden können.

Zur Interpretation des Spektrums bzw. der LC/MS-Features die durch eine Verbindung hervorgerufen werden, ist es daher notwendig viele in Frage kommende Peaks bzw. Features zu betrachten, in einen Zusammenhang zu stellen und daraus einen Schluss auf die Molekülmasse abzuleiten. Die Bestimmung der zu einem Molekül gehörenden Ionen wird gelegentlich, insbesondere im GC/MS Kontext, als *Dekonvolution* bezeichnet. Dieser Begriff ist jedoch nicht eindeutig definiert und wird im Bereich der Massenspektrometrie auch in anderen Zusammenhängen gebraucht, daher wird in dieser Arbeit der Begriff der *Annotation* benutzt.
Als *Annotation* von Peaks im Massenspektrum oder LC/MS-Features werden Verarbeitungsschritte bezeichnet, die Zusammenhänge zwischen verschiedenen Peaks/Features herstellen und einzelnen oder Gruppen von Peaks bzw. Features Informationen wie Ladungszustand, Ionenbezeichnung und insbesondere die Molekülmasse der gemessenen Substanz zuordnen. Weitergehende Annotationen wie die Zuweisung einer Summenformel oder die Identifizierung mittels einer Datenbanksuche sind möglich, werden aber hier nicht betrachtet.
Die dazu notwendigen Schritte bei der Verarbeitung eines Massenspektrums oder einer LC/MS-Feature-Liste sind

1. Gruppierung aller zu einem Molekül gehörenden Ionen
2. Bestimmung des Ladungszustandes dieser Ionen mithilfe der Isotopenpeaks
3. Bestimmung der Molekülmasse durch Interpretation der addukt- und fragmentspezifischen Massendifferenzen

Die Gruppierung lässt sich auch als Datenreduktion verstehen, da viele der in einem Massenspektrum erfassten Informationen redundant sind. Die Bestimmung der Molekülmasse ist der wichtigste und entscheidende Schritt der Annotation. Erst mit einer richtig bestimmten Molekülmasse sind weitere Verarbeitungsschritte wie die Dekomposition der Molekülmasse (Bestimmung einer möglichen Summenformel) und die Suche in Datenbanken möglich. Die hier behandelten Algorithmen zur Annotation wurden in [Tautenhahn et al. 2007] beschrieben und als R-Package implementiert (`http://msbi.ipb-halle.de/msbi/esi/`).

Verwandte Arbeiten

In [Görlach und Richmond 1999] wird ein unternehmensintern genutztes Programm mit graphischer Benutzeroberfläche zur interaktiven Ermittlung von Molekülmassen auf einzelnen Massenspektren beschrieben. Dabei wird eine fest implementierte Liste von Addukten genutzt, um zu einem vom Benutzer angewählten Peak im Spektrum über eine Massendifferenzbildung Addukte zu erkennen und darüber eine Molekülmasse vorzuschlagen. Dazu werden zunächst die monoisotopischen Peaks jedes Ions bestimmt und

die Intensität des ersten Isotopenpeaks mit einer Heuristik, ähnlich wie in Abschnitt 4.2.2 beschrieben, überprüft. Anschließend wird versucht, zunächst Quasi-Molekülionen wie $[M+H]^+$ zuzuordnen und anschließend Clusterionen wie $[2M+H]^+$, jeweils mit einer Toleranz von 0.2 *m/z*. Durch das Vorgehen in dieser Reihenfolge entstehen laut Aussage der Autoren keine Konflikte bei der Zuordnung. Zu dem angewählten Peak werden alle möglichen Positionen von Addukten im Spektrum markiert (auch für nicht vorhandene Peaks) und die berechnete Molekülmasse angezeigt.
Die Autoren von [Bayliss und Lashin 2006] beschreiben Merkmale der kommerziellen Software IntelliXtract (`http://www.acdlabs.com/products/spec_lab/exp_spectra/ms/intellixtract/features.html`). Die Bestimmung der zu einem Molekül gehörenden Ionen wird darin über einen Vergleich der Retentionszeiten der einzelnen Features durchgeführt, wobei die in einem Zeitfenster von 15-30% der Peakbreite bei halber Höhe (FWHM) liegenden Features als von einer Verbindung stammend zusammengruppiert werden. Weiterhin ist die Fähigkeit zur Erkennung der Adduktionen und insbesondere des $[M+H]^+$ Ions mittels eines „pattern matching approach“ erwähnt, aber nicht weiter erläutert.
Im Folgenden wird zunächst das Prinzip der regelbasierten Annotation dargestellt. Das Ziel dabei ist die Bestimmung der Molekülmasse durch Interpretation der beobachteten Addukte und Fragmente des Moleküls. Bestimmte Regeln für die Addukt- bzw. Fragmentbildung müssen dafür vorgegeben werden. Weiterhin wird eine korrelationsbasierte Methode vorgestellt, dessen Ziel die Gruppierung *aller* von einer Substanz hervorgerufenen Features ist. Beide Methoden können sowohl unabhängig voneinander eingesetzt, aber auch kombiniert werden.

4.2 Regelbasierte Annotation

Um Addukte und Fragmente eines Moleküls zu erkennen und damit die zugehörige Molekülmasse zu bestimmen, werden fest vorgegebene Regeln benutzt, die Informationen über die Massendifferenz und den Ladungszustand dieser Addukte bzw. Fragmente enthalten. Die im Folgenden beschriebene Methode wird daher – auch in Abgrenzung zu der in Abschnitt 4.3.2 beschriebenen korrelationsbasierten Methode – als *regelbasierte Annotation* bezeichnet.
Die regelbasierte Annotation kann sowohl auf einzelne Massenspektren (z.B. FT-ICR Spektren) als auch auf Feature-Listen von LC/MS Daten angewendet werden. Im Weiteren wird jedoch nur auf die Annotation von LC/MS-Features eingegangen.

4.2.1 Zeitliche Gruppierung

Um die in einer Liste erfassten LC/MS-Features zu annotieren, ist zunächst eine zeitliche Gruppierung notwendig, um potentiell zusammengehöriger Features zu erfassen, welche

dann weiter untersucht werden können.
Während eine Substanz M eluiert, treten die bei der Ionisierung dieser Substanz gebildeten Ionen in annähernd gleichem Verhältnis auf. Die chromatographischen Peaks für die einzelnen Ionen haben daher die gleiche Form, insbesondere tritt das Peakmaximum zum exakt selben Zeitpunkt auf (siehe z.B. Abbildung 4.1 auf Seite 76). Theoretisch sollten daher alle Features die zu einer Substanz gehören, auch mit exakt derselben Zeitkoordinate erfasst werden. In der Praxis ist jedoch zu beobachten, dass insbesondere für Ionen die mit geringer Intensität auftreten, der chromatographische Peak eher klein und flach ausfällt, die Peakform mehr vom Rauschen beeinflusst wird und daher das Peakmaximum schwerer zu bestimmen ist. In Folge dessen kann der bei der Feature-Detektion erfasste zeitliche Mittelpunkt der Features dieser Ionen leicht verschieden sein. Diese Abweichungen sind beispielsweise in der Tabelle 8.17 im Anhang (Seite 127) zu beobachten.
Aus diesem Grund wird zur zeitlichen Gruppierung keine feste Abgrenzung benutzt, sondern ein gleitendes Fenster über die Retentionszeit. Die Fensterbreite wird in Abhängigkeit von der verwendeten Chromatographie gewählt, welche die Breite der chromatographischen Peaks bestimmt. Alle in einem solchen Retentionszeitfenster liegenden Features werden in den folgenden Schritten auf Zusammenhänge untersucht und annotiert. Die durch die zeitliche Überlappung eventuell auftretenden redundanten Annotationen werden in einem Folgeschritt wieder entfernt.

4.2.2 Isotopomere

Wie schon in Abschnitt 2.3 erwähnt, können neben dem monoisotopischen Peak[9] weitere Isotopenpeaks entsprechend der auftretenden Isotopomere beobachtet werden.
Dabei unterscheiden sich die Isotopomere eines Moleküls bezüglich der Kombination von Isotopenspezies im Molekül, beispielsweise kann für jedes Kohlenstoffatom entweder ^{12}C oder ^{13}C „verbaut“ worden sein. Ein hypothetisches Molekül (es bestehe nur aus Kohlenstoff), in dem ein Kohlenstoffatom durch ^{13}C ersetzt wurde, und alle anderen aus ^{12}C bestehen, unterscheidet sich von seinem Isotopomer, in dem alle Kohlenstoffatome ^{12}C sind, nur durch ein Neutron. Der Massenunterschied beträgt jedoch nicht eine Neutronenmasse ($m_n = 1.008665u$), sondern deutlich weniger ($1.003355u$). Grund ist der sogenannte *Massendefekt* (auch *Massenverlust*), die auftretende Differenz der Masse eines Atoms gegenüber der Summe der Ruhemassen seiner Kernbausteine, was als relativistischer Effekt mit der aufgewendeten Bindungsenergie erklärt wird [Gross 2004].
Die Differenz zwischen Isotopomeren, die sich nur in Bezug auf ein Stickstoffatom unterscheiden, ^{15}N statt ^{14}N, beträgt $0.997u$. Ein solcher ^{15}N-Peak würde sich also vom ^{13}C-Peak um nur 0.006355 u unterscheiden. Um diese beiden Peaks bei einer Molekülmasse von

[9] Die Erläuterungen in diesem Abschnitt beziehen sich aufgrund der chem. Begriffskonventionen auf Peaks eines Spektrums, die Zusammenhänge gelten aber genauso für die entsprechenden Features.

z.B. 500 getrennt darzustellen, wäre ein Massenspektrometer mit einer Auflösung von $R \geq 78678$ notwendig ($R = m/\Delta m = 500/0.006355 \approx 78678$). Solch hohe Auflösungen werden derzeit nur von wenigen (kostenintensiven) Geräten wie z.B. FT-ICR-Massenspektrometern erreicht. Das TOF-Massenspektrometer (Bruker microTOF-Q), mit dem mehrere in dieser Arbeit verwendeten Datensätze gemessen wurden, besitzt eine Auflösung von ca. 14000, die beschriebenen Peaks werden daher nicht getrennt dargestellt. In gleicher Art und Weise überlagern sich in den Messungen auch die Isotopenpeaks anderer Elemente, wie z.B. H, O, S.

Aus diesen Gründen ist der beobachtete Abstand zwischen Isotopenpeaks variabel. Bei den in dieser Arbeit betrachteten kleinen Molekülen, gemessen im Bereich von 100 bis 1000 m/z, liegt er bei ca. 1.001 bis 1.0035 u. Diese Abschätzung basiert auf berechneten Isotopenmustern für 702 Substanzen aus der AraCyc Datenbank [Mueller et al. 2003], durchgeführt von Carsten Kuhl, IPB Halle. Für den Abstand zwischen monoisotopischem Peak und erstem Isotopenpeak („^{13}C-Peak“) beträgt der Median aus diesen berechneten Werten 1.0032, zwischen erstem und zweitem Isotopenpeak 1.002, zwischen zweitem und drittem Isotopenpeak 1.0027.

Da die Bildung eines geeigneten Modells hierfür (ähnlich dem für Proteine bzw. Peptide benutzten Averagine-Modell [Senko et al. 1995]) noch aussteht, wird als Anhaltspunkt für den Abstand benachbarter Isotopenpeaks $\widetilde{m}_n$ momentan ein Wert von $\widetilde{m}_n = 1.0025u$ genutzt, in Verbindung mit einer ausreichend groß gewählten Toleranz, für die auch die Genauigkeit des Massenspektrometers zu beachten ist.

Für mehrfach geladene Ionen I^z gilt, dass der Abstand benachbarter Isotopenpeaks entsprechend $\approx \widetilde{m}_n/z$ beträgt. Dementsprechend lässt sich aus den beobachteten Abständen der Isotopenpeaks eines Ions auf dessen Ladungszustand schließen. Die Annotation der Isotopenpeaks erfolgt durch die Suche nach Mustern in den m/z-Werten der Features welche im Abstand von

$$\Delta m/z = n\frac{\widetilde{m}_n}{z}\ ,\ z = 1, 2, \ldots, Z,\ n = 1, 2, \ldots, N \tag{4.2}$$

zueinander stehen. Für n und z werden für die Suche Obergrenzen N und Z vorgegeben. Bei den hier betrachteten LC/MS-Datensätzen ist der höchste (und selten) vorkommende Ladungszustand drei, häufiger treten einfach und zweifach geladene Ionen auf, Z kann daher auf 3 gesetzt werden. Beobachtet werden häufig der erste und der zweite Isotopenpeak, seltener auch der dritte und nur bei entsprechend hoher Intensität auch der vierte Isotopenpeak. Die Suche kann daher mit N=4 beschränkt werden.

Über die Abstände der gefundenen Isotopenmuster lässt sich nun direkt auf die Ladung z des Ions rückschließen, wofür ein erkannter Isotopenpeak ausreicht. Nach der Zuordnung der Isotopenpeaks sind für die meisten Ionen nach diesem Schritt der Ladungszustand und damit die monoisotopische Masse bekannt und können für die weiteren Berechnungen verwendet werden. Eine Ausnahme bilden Ionen, die nur mit sehr geringer Intensität

auftreten. Dabei kommt es vor, dass neben dem monoisotopischen Peak kein weiterer Isotopenpeak beobachtet werden kann. In solchen Fällen kann kein Ladungszustand erkannt werden. Treten unterschiedliche Abstände bei den Isotopenpeaks auf, so deutet dies auf eine Überlagerung von Isotopenmustern verschiedener Substanzen hin. Auch in diesem Fall kann kein Ladungszustand zugeordnet werden. Dementsprechend steht für solche Ionen die Information über den Ladungszustand für die im nächsten Abschnitt beschriebene Erkennung der Addukte und Fragmente nicht zur Verfügung. Bezüglich des Ladungszustandes konkurrierende Annotationen können dann in einigen Fällen nicht aufgelöst werden.

Um die korrekte Zuordnung der erkannten Isotopenpeaks zu überprüfen, lässt sich zumindest für den ersten Isotopenpeak, dessen Intensität bei den hier betrachteten, kleinen organischen Molekülen durch ^{13}C dominiert wird, ein einfacher Test durchführen:
Die natürliche Häufigkeit des ^{12}C-Isotops beträgt 98.9 % ($p_{^{12}\mathrm{C}} = 0.989$), die des ^{13}C-Isotops ist 1.1% ($p_{^{13}\mathrm{C}} = 0.011$). Die Wahrscheinlichkeit, bei einem zufällig gewählten Molekül mit insgesamt $\mathrm{N_C}$ Kohlenstoffatomen eines zu erhalten, dass nur ^{12}C enthält, ist damit $P_{\mathrm{N_C}\,^{12}\mathrm{C}} = (p_{^{12}\mathrm{C}})^{\mathrm{N_C}}$. Die Wahrscheinlichkeit ein Molekül zu erhalten, in dem eines dieser $\mathrm{N_C}$ Kohlenstoffatome ein ^{13}C Atom ist, beträgt $P_{1\,^{13}\mathrm{C}} = p_{^{13}\mathrm{C}}(p_{^{12}\mathrm{C}})^{\mathrm{N_C}-1}$. Da es bei einem Molekül mit $\mathrm{N_C}$ Kohlenstoffatomen $\mathrm{N_C}$-mal die Möglichkeit gibt, ein Molekül mit einem ^{13}C Atom zu erhalten, ergibt sich das Intensitätsverhältnis $\mathrm{RI}_{^{13}\mathrm{C}/\,^{12}\mathrm{C}}$ von ^{13}C zu ^{12}C Isotopenpeak als

$$\mathrm{RI}_{^{13}\mathrm{C}/\,^{12}\mathrm{C}} = \mathrm{N_C} \cdot \frac{p_{^{13}\mathrm{C}}(p_{^{12}\mathrm{C}})^{\mathrm{N_C}-1}}{(p_{^{12}\mathrm{C}})^{\mathrm{N_C}}} = \mathrm{N_C} \cdot \frac{p_{^{13}\mathrm{C}}}{p_{^{12}\mathrm{C}}} = \mathrm{N_C} \cdot \frac{0.011}{0.989} = \mathrm{N_C} \cdot 0.01112235 \approx \mathrm{N_C} \cdot 0.011.$$

Zu beachten ist, dass auch noch andere Isotope zur Intensität des ersten Isotopenpeaks beitragen, beispielsweise ^{15}N mit einer Häufigkeit von 0.366% oder ^{2}H mit einer Häufigkeit von 0.015%. Da, wie schon erwähnt, für Metabolite kein geeignetes Modell für die Zusammensetzung vorhanden ist und die Intensität des ersten Isotopenpeaks bei den hier betrachteten Molekülen hauptsächlich durch ^{13}C bestimmt wird, kann die Abschätzung des Intensitätsverhältnisses nur über die Anzahl der Kohlenstoffatome im Molekül vorgenommen werden. Das Verhältnis $\mathrm{RI}_{^{13}\mathrm{C}/\,^{12}\mathrm{C}}$ wird damit als Beschreibung der Obergrenze des Intensitätsverhältnis von ^{13}C zu ^{12}C-Peak genutzt, die bei einer Molekülzusammensetzung aus {C,H,N,O,S,P} nur dann erreicht werden kann, wenn das Molekül ausschließlich aus Kohlenstoff besteht. Die Anzahl der Kohlenstoffatome im Molekül kann nach oben abgeschätzt werden, indem angenommen wird, dass das Molekül nur aus Kohlenstoffatomen besteht. Dazu wird die gemessene Masse m des Ions durch die Masse von Kohlenstoff ($= 12\ u$) geteilt. Der erhaltene Wert $\widetilde{N_C} = \left\lfloor \frac{m}{12} \right\rfloor$ wird für die Überprüfung verwendet, ob

$$\text{Intensität}(^{13}\mathrm{C}) < \text{Intensität}(^{12}\mathrm{C}) \cdot \widetilde{N_C} \cdot \mathrm{RI}_{^{13}\mathrm{C}/\,^{12}\mathrm{C}}$$

Ist dies nicht der Fall, besteht Grund zur Annahme, dass die Zuordnung der Isotopenpeaks in diesem Fall nicht korrekt ist, oder aber eine Überlagerung mit einem anderen

Massensignal vorliegt[10]. Die Annotation des Isotopenmusters wird in diesem Fall verworfen.

4.2.3 Addukte und Fragmente

Die Suche nach Addukten und Fragmenten wird ebenfalls unter Benutzung vorgegebener Differenzen durchgeführt. Vom Benutzer wird dazu eine Tabelle mit häufig auftretenden Addukten und Fragmenten vorgegeben, mit der Masse der chemischen Modifikation und weiteren Angaben zum Ion.

$$s = \frac{n\ \mathrm{mass}(M) + \overbrace{\mathrm{mass}(\{I\}) + \mathrm{mass}(\{L\}) - \mathrm{mass}(\{F\})}^{x}}{z} \tag{4.3}$$

Die Masse x der chemischen Modifikation (siehe auch Beschreibung auf Seite 65) wird dabei für angelagerte Ionen, aus dem Lösungsmittel stammenden Molekülen und abgespaltene Fragmente zusammen angegeben. Gleichfalls wird die Ladung z des Ions als aus allen am Bildungsprozeß beteiligten Ionen resultierender Ladungszustand angegeben. Tabelle 4.2 zeigt ein Beispiel für solche Ionisierungsregeln. Eine ausführlichere Liste von Regeln, die auch für die später beschriebenen Versuche benutzt wurde, findet sich im Anhang auf Seite 124.

Index	Ion	x: Masse der chemischen Modifikation in [u]	n: Anzahl der Moleküle M im Ion	z: Ladungszustand
1	$[M+H]^+$	1.007276	1	1
2	$[M+Na]^+$	22.98922	1	1
3	$[M+H+Na]^{2+}$	23.9965	1	2
4	$[M+K]^+$	38.96316	1	1
5	$[2M+H]^+$	1.01	2	1
6	$[2M+Na]^+$	22.98922	2	1
7	$[M+H\text{-}NH_3]^+$	-16.01817	1	1
8	$[M\text{-}C_3H_9N]^+$	-59.07295	1	1

Tabelle 4.2: Beispiele für bekannte Addukte und Fragmente im positiven Ionisierungsmodus mit der Masse der chemischen Modifikation. Die tatsächlichen Differenzen in m/z für eine mutmaßliche Molekülmasse m=mass(M) werden unter Beachtung des Ladungszustandes und der Anzahl der Moleküle M im Ion berechnet.

[10]Davon abgesehen ist zu beobachten, dass bei Peaks mit geringer Intensität oder aber sehr hoher Intensität im Bereich der Detektorsättigung das Isotopenmuster verfälscht sein kann.

Da jede Ionisierungsregel bereits die Gesamtmasse der chemischen Modifikation, sowie die Gesamtladung z beschreibt, lässt sich das Massensignal des Ions in Vereinfachung von Gleichung 4.1 beschreiben als:

$$s = \frac{nm + x}{z} \tag{4.4}$$

wobei s das beobachtete Massensignal ($[m/z\ ]$) des Ions ist, gebildet aus n Molekülen der Masse m, einer chemischen Modifikation mit der Masse x und der Ladung z.
Die Umstellung von Gleichung (4.4) nach m ergibt

$$m = \frac{zs - x}{n} \tag{4.5}$$

und beschreibt die Rückrechnung der Molekülmasse bezüglich eines Massensignals s und bekannter Ionisierungsregel $\{x, n, z\}$.
Für jedes Retentionszeitfenster werden die möglichen Kombinationen

$$m_{i,j} = \frac{z_j s_i - x_j}{n_j} \tag{4.6}$$

von Massensignalen $s_i\ ,\ i = 1, \ldots, N$ im Zeitfenster und den vorgegebenen Ionisierungsregeln $X_j = (n_j, z_j, x_j)\ ,\ j = 1, \ldots, J$ betrachtet. Wurde bei dem in Abschnitt 4.2.2 beschriebenen Schritt der Ladungszustand von s_i erkannt, so werden dabei nur die Regeln X_j benutzt, bei denen die Ladung z_j der von s_i entspricht.
Jede Gruppe von errechneten Massen $m_{i,j}$, die bis auf eine vorgegebene Toleranz den gleichen Wert haben, bildet eine *Annotations-Hypothese*. Dazu wird die Differenzmatrix $m_{i,j}$ in einen Vektor mit ansteigend sortierten Werten m umgewandelt, wobei die ursprünglichen Indizes i, j zusätzlich gespeichert werden. In diesem Vektor werden jeweils zwei benachbarte Werte daraufhin untersucht, ob die Differenz dieser Werte kleiner als ein vom Benutzer vorgegebener Schwellwert[11] ist. Jede Gruppe von aufeinanderfolgenden Werten, die diese Bedingung erfüllen, bilden eine Annotations-Hypothese. In einer Matrix $m_{i,j}$ können daher mehrere solche Gruppen gefunden werden, wobei jede Gruppe aus mindestens zwei Werten besteht. Die Indizes $\{(i, j)\}$ jeder Gruppe werden als eine Annotations-Hypothese gespeichert und beschreiben somit eine hypothetische Molekülmasse $\overline{\mathrm{m}}$ zu einem Molekül M, von dem die Addukte bzw. Fragmente $\{X_j\}$ beobachtet wurden.

Beispiel:
Innerhalb eines Retentionszeitfensters werden die Massensignale $(s_1, s_2, s_3, s_4) = (101.107,$ 123.088, 139.064, 142.118) m/z beobachtetet. Unter Benutzung der Gleichung 4.6 in Ver-

[11] Die zulässige Toleranz wird in Abhängigkeit von der Genauigkeit des Massenspektrometers gewählt, die Voreinstellung liegt bei 0.01 m/z.

bindung mit den Regeln in Tabelle 4.2 ergeben sich unter anderem die Kombinationen

$$m_{1,1} = \frac{z_1 s_1 - x_1}{n_1} = \frac{1 \cdot 101.107 - 1.007276}{1} = 100.0997$$
$$m_{2,2} = \frac{z_2 s_2 - x_2}{n_2} = \frac{1 \cdot 123.088 - 22.98922}{1} = 100.0988$$
$$m_{3,4} = \frac{z_4 s_3 - x_4}{n_4} = \frac{1 \cdot 139.064 - 38.96316}{1} = 100.1008$$

Bei einer vorgegebenen Toleranz von 0.01 *m/z* werden die Werte $m_{1,1}$, $m_{2,2}$ und $m_{3,4}$ als ähnlich betrachtet und bilden die Annotations-Hypothese, dass s_1, s_2 und s_3 die Addukte $[M+H]^+$,$[M+Na]^+$ und $[M+K]^+$ eines Moleküls M mit einer (durch Mittelung von $m_{1,1}$, $m_{2,2}$ und $m_{3,4}$) geschätzten Masse $\overline{m}$ von $100.0998u$ sind. Für das Massensignal s_4 ergab sich mit den vorgegebenen Regeln keine sinnvolle Differenz, dieses wird daher außer Acht gelassen. Als Annotations-Hypothese H_1 gespeichert werden in diesem Beispiel die Indizes $H_1 = \{(i,j)\} = \{(1,1),(2,2),(3,4)\}$.

Wenn vom Benutzer diese Option gewählt wurde, wird direkt nach der Hypothesenbildung eine korrelationsbasierte Verifikation (Genaueres im nächsten Abschnitt) durchgeführt. Dabei wird für jedes Paar von Features innerhalb einer Annotations-Hypothese überprüft, ob deren chromatographische Intensitäten korrelieren. Wenn diese Bedingung für ein Feature k für mindestens eine Paarbildung erfüllt ist, gilt dessen Zugehörigkeit als verifiziert. Andernfalls werden die Indizes $(i,j), i = k$ aus der Annotations-Hypothese entfernt.
Da die Retentionszeitfenster, in denen die Annotationen durchgeführt werden, über ein gleitendes Fenster gebildet werden, entstehen mehrfache und zum Teil konkurrierende Annotations-Hypothesen bezüglich derselben Massensignale. Um diese Konflikte zu beseitigen, werden zunächst solche Hypothesen betrachtet, deren Schnittmenge bezüglich der Indizes i nicht leer ist, die also Annotationen in Bezug auf dieselben Massensignale enthalten. Wenn eine Annotations-Hypothese in Bezug auf die Indizes i und j Teilmenge einer anderen Hypothese ist, so kann sie entfernt werden. In anderen Fällen ist die Konfliktauflösung schwieriger. Enthalten zwei unterschiedliche Annotations-Hypothesen eine nichtleere Schnittmenge von Indizes i und j, so ist schwer zu entscheiden, welche der Hypothesen die zutreffende ist. Bei Hypothesen mit unterschiedlicher Kardinalität $|H_k|$ ist diejenige mit der größeren Kardinalität zu bevorzugen, da diese Hypothese gewissermaßen mehr Argumente besitzt. Im angeführten Beispiel wäre z.B. auch die Hypothese $H_2 = \{(i,j)\} = \{(1,5),(2,6)\}$ ($[2M+H]^+$, $[2M+Na]^+$) errechnet worden, mit $\overline{m} = 50.04963u$. Hier ist die Kardinalität von $|H_1| = 3$ größer als die von $|H_2| = 2$.
Bei gleicher Kardinalität wird der Konflikt nicht aufgelöst, sondern beide Hypothesen annotiert. Ein Beispiel dafür sind die konkurrierenden Hypothesen ($[M+H]^+$, $[M+Na]^+$) sowie ($[2M+H]^+$, $[2M+Na]^+$), die dann beobachtet werden, wenn nur diese eine Massendifferenz auftritt und nicht mindestens ein weiteres Ion diesen Konflikt auflöst. Die erstere

Hypothese wäre jedoch aus chemischer Sicht zu bevorzugen, da das Auftreten von Clusterionen wie $[2M+H]^+$, ohne dass auch $[M+H]^+$ beobachtet wird, unwahrscheinlich ist. Die durch den Benutzer frei editierbare Regeltabelle bietet den Vorteil der leichten Anpassbarkeit für andere Geräte bzw. Ionisierungsprozesse, sowie die Möglichkeit, auch Regeln für substanzklassenspezifische Fragmente, wie z.B. $[M\text{-}C_3H_9N]^+$ für Cholinester, einzugeben. Ein Nachteil ist jedoch, dass weiterführende Regeln zur Konfliktauflösung, wie die erwähnte Bevorzugung von Quasi-Molekülionen gegenüber Clusterionen, damit schwer zu implementieren sind. Eine mögliche Lösung für dieses Problem wäre, die Addukt/Fragment-Regeln nicht mehr wie in der hier beschriebenen Art vom Benutzer vorgegeben zu lassen, sondern programmintern in einer Schleife über viele mögliche Modifikationen (in der Art von Gleichung 4.1) zu generieren, was eine Formalisierung bezüglich der Ionenarten erleichtern würde.
Wie im Abschnitt 4.2.2 erwähnt, kommt es vor, dass für einige Ionen kein Ladungszustand erkannt wird. In solchen Fällen ergeben sich gelegentlich Konflikte, die nicht aufgelöst werden. Werden beispielsweise zwei Ionen m/z 139 und 51 beobachtet und nur für das erstere ist der Ladungszustand (=+1) bekannt, so ergeben sich je nach verwendeter Regelmenge mehrere Möglichkeiten zur Annotation (nur gerundete Werte): Die Annotation $[M+K]^+$ und $[M+2H]^{2+}$ erklärt die Differenz $(100+39)$ - $\frac{100+2}{2} = 88$ für eine Molekülmasse M von $100u$, aber ebenfalls $[2M+K]^+$ und $[M+H]^+$ mit der Differenz $(2 \cdot 50 + 39)$ - (50+1) = 88 für eine Molekülmasse M von $50u$. Aufgrund des unbekannten Ladungszustandes für das zweite Ion kann ohne die Beobachtung eines weiteren Addukts ($\rightarrow$ Kardinalitätsregel) keine dieser beiden Annotations-Hypothesen verworfen werden. In derartigen Fällen werden alle sich ergebenden Möglichkeiten zusammen mit der geschätzten Molekülmasse als konkurrierende Annotations-Hypothesen annotiert.

4.3 Korrelationsanalyse der Chromatogramme

Wie schon erwähnt, treten in dem Zeitraum, in dem eine Substanz M eluiert, die bei der Ionisierung dieser Substanz gebildeten Ionen in annähernd gleichem Verhältnis auf. Die beobachteten chromatographischen Peaks für die einzelnen Ionen haben daher die gleiche Form, jedoch unterschiedliche Höhe, entsprechend der Signalintensität für die einzelnen Ionen. Betrachtet man den Intensitätsverlauf über die Zeit von zwei aus einem Molekül M gebildeten Ionen, so zeigt sich aufgrund des konstanten Bildungsverhältnisses der Ionen ein linearer Zusammenhang. Dieser Zusammenhang ist um so ausgeprägter, je größer die Intensitäten der einzelnen Ionen sind (Abbildung 4.1). Chromatographische Peaks mit geringer Intensität sind eher vom Rauschen beeinflusst, die Peakform fällt deutlich flacher und eher „gezackt“ aus, der lineare Zusammenhang ist daher geringer. Der Grad des linearen Zusammenhangs der chromatographischen Intensitäten zweier Ionen I_X und I_Y lässt sich mittels des empirischen Pearson-Korrelationskoeffizienten r über die gemessenen

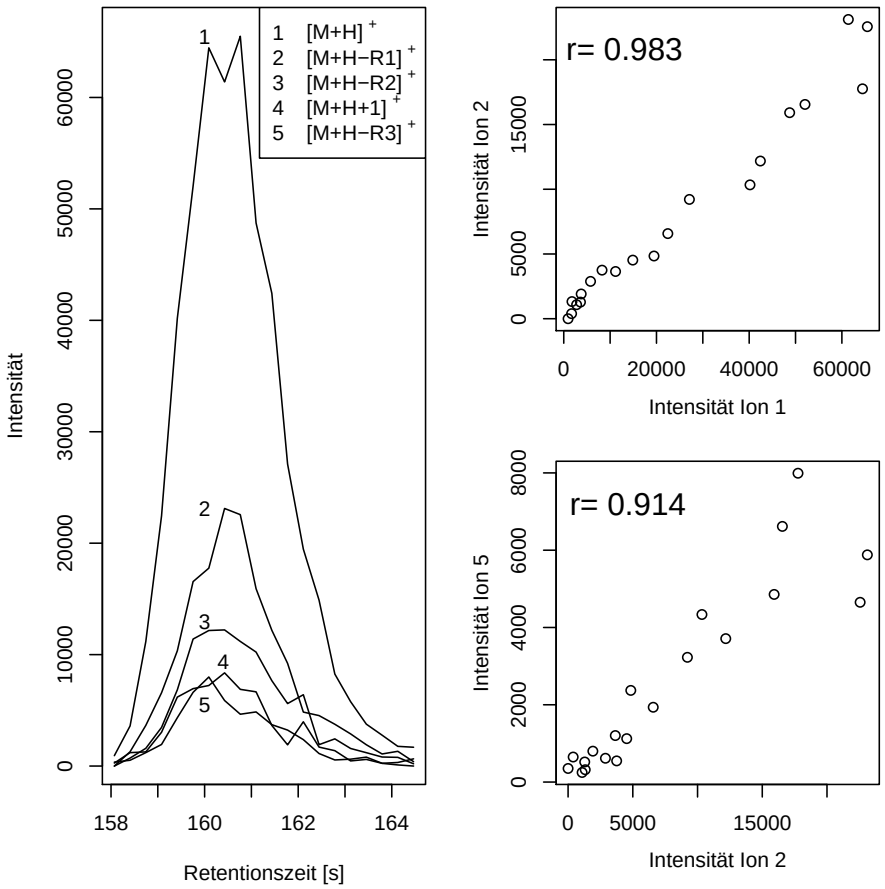

Abbildung 4.1: Intensitätskorrelationen. Links: Extrahierte Ionen Chromatogramme von fünf ausgewählten Ionen von Kinetin (Datensatz MM14). R1=$C_5H_5N_5$, R2=C_4H_4O, R3=CO. $[M+H+1]^+$ bezeichnet den ersten Isotopenpeak von $[M+H]^+$. Rechts: Gegeneinander aufgetragene Intensitätswerte zweier ausgewählter Ionenpaare. Angegeben ist der empirische Pearson-Korrelationskoeffizient r.

Intensitäten X und Y der Länge n schätzen:

$$\hat{\varrho}(X,Y) := r_{XY} := \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\,\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

Um die positive Korrelation nachzuweisen, stellt man die Hypothesen $H_0 : \varrho(X,Y) \leq 0$ und $H_1 : \varrho(X,Y) > 0$) auf und berechnet die Testgröße

$$T(X,Y) = \frac{r_{XY}\sqrt{n-2}}{\sqrt{1-(r_{XY})^2}}$$

die unter der Modellannahme der Normalverteilung von X und Y eine t-Verteilung mit $n-2$ Freiheitsgraden besitzt. Sei $t_{n-2}(p)$ das p-Quantil der t-Verteilung mit $n-2$ Freiheitsgraden und α die gegebene Irrtumswahrscheinlichkeit erster Art, dann wird H_0 abgelehnt,

falls $|T| \geq t_{n-2}(1-\alpha)$ [Grabowski 2004]. Verwendet wird eine Implementierung in der R-Funktion *cor.test*, bei einer Irrtumswahrscheinlichkeit α von 0.05.
Der Test unter Berücksichtigung der Anzahl der Messwerte n ist insbesondere dann von Bedeutung, wenn für Peaks mit niedrigen chromatographischen Intensitäten nur sehr wenige Messwerte vorliegen, da bei der Centroidisierung alle Datenpunkte im Spektrum unterhalb eines bestimmten Schwellwertes unterdrückt werden. Unter Umständen wird dann für derartige Peaks, bei denen nur drei oder vier gemeinsame Messwerte vorliegen, zwar ein empirischer Korrelationskoeffizient über dem vorgegebenen Schwellwert beobachtet, die Korrelation aber als nicht signifikant abgelehnt.
In den durchgeführten Versuchen wurde unter Verwendung des Tests eine Verringerung der falsch positiven Korrelationen bei Peaks niedriger Intensität festgestellt. Das Risiko von falsch Negativen, d.h. dass zu einer Substanz gehörende Peaks fälschlicherweise als unkorreliert angenommen werden, wird im Sinne einer möglichst sicheren Zuordnung in Kauf genommen.
Implementiert wurden zwei Möglichkeiten der korrelationsbasierten Analyse, die sich bezüglich Ergebnis und Laufzeit stark unterscheiden. Die Variante mit der geringsten Laufzeit besteht darin, zunächst die regelbasierte Annotation anzuwenden und darauffolgend eine Verifikation der Annotation durch Korrelationsanalyse (siehe nächster Abschnitt) durchzuführen. Bei dieser Variante beträgt die Laufzeit auf einem *A.thaliana* Blattextrakt aus dem Datensatz LSMIX (siehe Anhang) mit 5166 detektierten Features 4.5 Minuten auf einer 2.4 GHz CPU.
Bei der zweiten Variante wird zunächst eine korrelationsbasierte Gruppierung durchgeführt und auf den daraus erhaltenen Feature-Gruppen die regelbasierte Annotation durchgeführt. Der Vorteil hierbei ist eine umfassende Gruppierung aller von einer Substanz hervorgerufenen Features (mehr im übernächsten Abschnitt), welcher jedoch mit einer erhöhten Laufzeit erkauft wird. So beträgt die Gesamtlaufzeit auf derselben Messung bei dieser Variante 19 Minuten, wovon rund 15 Minuten auf die korrelationsbasierte Gruppierung entfallen.

4.3.1 Verifikation der Annotation durch Korrelationsanalyse

Um zu überprüfen, dass die durch die regelbasierte Annotation zugeordneten Features auch tatsächlich in Zusammenhang stehen, kann aufgrund des erwähnten linearen Zusammenhangs der Intensitäten eine Korrelationsanalyse der entsprechenden Chromatogramme durchgeführt werden. Dazu wird für jedes Feature-Paar innerhalb einer Annotations-Hypothese der Korrelationskoeffizient für die chromatographischen Intensitäten der Ionen berechnet und der Test auf Signifikanz der Korrelation durchgeführt. Wenn die Korrelationsbedingung für ein Feature für mindestens eine Paarbildung erfüllt ist, gilt dessen Zugehörigkeit zu der Annotations-Hypothese als verifiziert. Andernfalls wird dieses Fea-

ture aus der Annotations-Hypothese entfernt.
Bei dieser Methode werden somit lediglich Aussagen zu den bereits annotierten Features erhalten, alle anderen Features werden nicht betrachtet.

4.3.2 Korrelationsbasierte Gruppierung

Durch die regelbasierte Annotation können nur solche Zusammenhänge gefunden werden, für die aufgrund von Erfahrungswerten für die Ionisierungsvorgänge und Fragmentbildungen Regeln vorgegeben wurden. Da die bei der Ionisierung und kollisionsinduzierten Dissoziation ablaufenden Prozesse hochkomplex sind, ist es praktisch unmöglich, eine Regelmenge zu erstellen, die *alle* Möglichkeiten der Ionenbildung aus einem Molekül erfasst. In der Praxis ist es jedoch hilfreich zu wissen, welche Ionen – neben denen die durch die Regelanwendung annotiert wurden – zum selben Molekül gehören, auch wenn für diese dann außer dem Fakt des Zusammenhangs keine weiteren Informationen angegeben werden können. Eine Gruppierung von allen Ionen eines Moleküls erleichtert die nachfolgende Interpretation und ermöglicht (über die Anzahl der Gruppen) eine Abschätzung, wieviele Substanzen tatsächlich in der Probe gemessen wurden. Für LC/MS-Features kann eine solche Gruppierung (zusätzlich oder unabhängig von der regelbasierten Annotation) durch Ausnutzung der beschriebenen korrelativen Zusammenhänge zwischen den gemessenen Intensitäten der Ionen eines Moleküls durchgeführt werden.
Verwendet wird dazu wiederum ein gleitendes Fenster über die Retentionszeit, nur werden jetzt im Gegensatz zur im vorherigen Abschnitt beschriebenen Verifikation der annotierten Features *alle* Features innerhalb der Retentionszeitfensters miteinander korreliert. Alle möglichen Paarbildungen von Features werden betrachtet und der Pearson-Korrelationskoeffizient sowie der Korrelationstest für die chromatographischen Intensitäten dieses Paares berechnet. Liegt der Korrelationskoeffizient über dem vorgegebenen Schwellwert, so wird dieses Feature-Paar in einer Liste gespeichert.
Aus der Liste der korrelierenden Feature-Paare kann nun ein ungerichteter Graph G gebildet werden, wobei jedes Feature einen Knoten bildet und zwischen den korrelierenden Features eine Kante gezogen wird. Unter optimalen Voraussetzungen (gute chromatographische Trennung der Substanzen, ausreichend hohe Intensitäten der Features, optimale Wahl des Schwellwertes) bilden alle Features die zu einer Substanz gehören, eine Zusammenhangskomponente mit hoher Kantenkonnektivität in G. In der Praxis zeigen Features mit geringer Intensität jedoch zum Teil nur sehr schwache Korrelationen, diese sind daher zu der Zusammenhangskomponente der Substanz nur schwach verbunden, oder sind, da der Korrelationskoeffizient unterhalb des Schwellwertes lag, gar nicht in der Zusammenhangskomponente enthalten. Wird der Schwellwert herabgesetzt, um dieses Problem zu vermeiden, führt dies vermehrt dazu, dass Kanten zwischen Features auftreten, die zu verschiedenen, koeluierenden Substanzen gehören. Abbildung 4.2 (links) zeigt ein solches

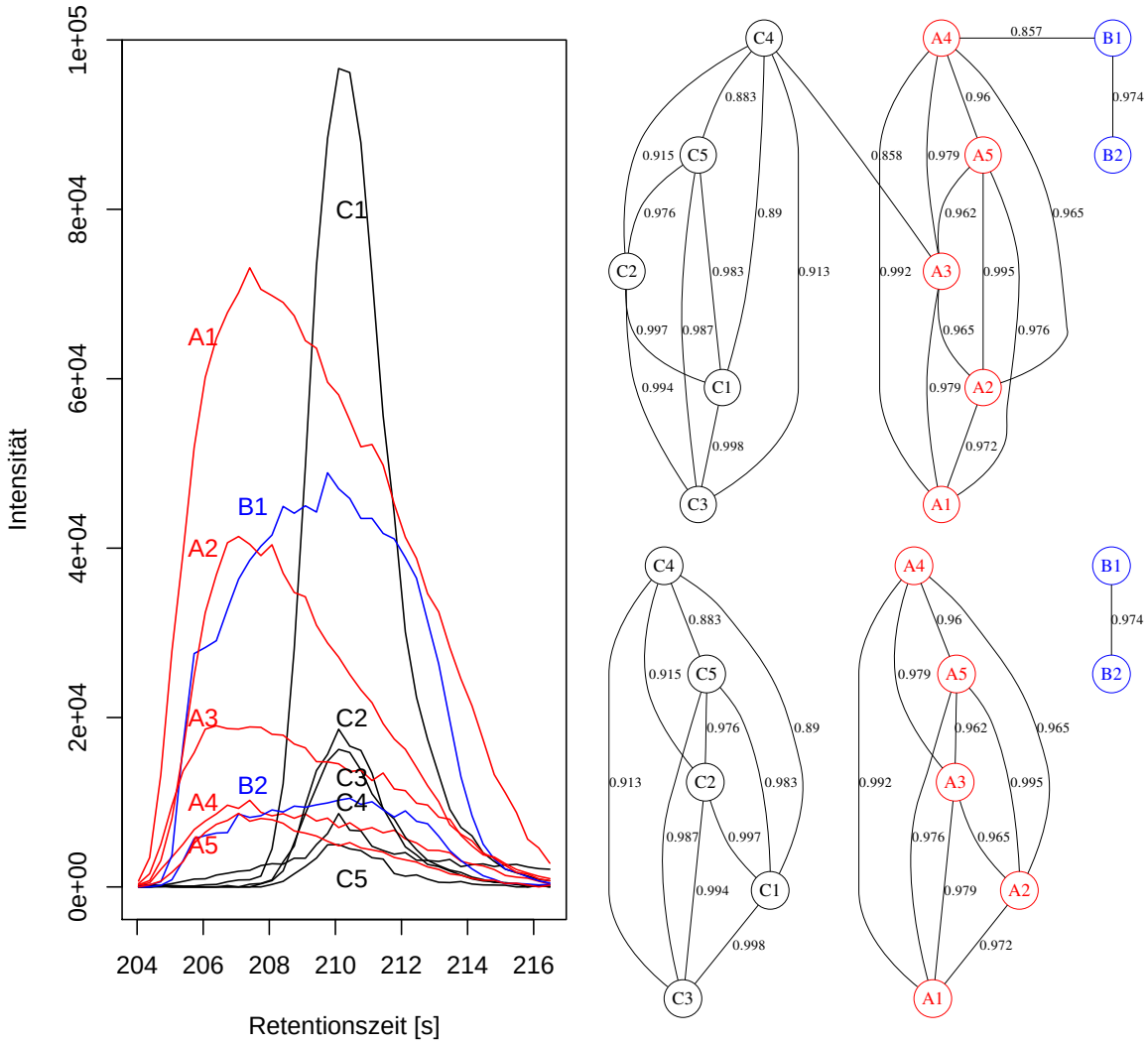

Abbildung 4.2: Intensitätskorrelationen bei koeluierenden Substanzen (aus Datensatz E170, Details im Anhang). Links: Extrahierte Ionen Chromatogramme von ausgewählten Ionen der Substanzen Methoxy-indolyl-3-methylglucosinulat(A), Coumaroylagmatin(C) sowie einer unbekannten Substanz(B). Rechts oben: aus den Intensitätskorrelationen aufgebauter Graph. Rechts unten: Graph nach Zerlegung in stark verbundene Subgraphen. Die als Kantengewichte angegebenen Korrelationskoeffizienten spielen bei der Graphzerlegung keine Rolle. Die Breite des gleitenden Fenster für die Auswahl der Features wurde mit 2 s gewählt, der Schwellwert für den Korrelationskoeffizienten betrug 0.85.

Beispiel für drei koeluierende Substanzen. In derselben Abbildung rechts oben ist der aus der Liste aller korrelierenden Feature-Paare aufgebaute Graph gezeigt. Die Einfärbungen und Bezeichungen im Graphen dienen dabei ausschließlich der Visualisierung, dem Algorithmus sind diese nicht bekannt. Die Intensitätskorrelationen zwischen allen Features einer Substanz sind deutlich ausgeprägt, was sich an den hohen Korrelationskoeffizienten und damit starken Kantenkonnektivität im Graphen zeigt. Jedoch sind aufgrund der

Erfassung der Features in einem gleitenden Fenster und der (etwas schwächeren) Korrelation von chemisch nicht zusammengehörigen Feature-Paaren auch solche Kanten vorhanden (C4-A3 und A4-B1), die bewirken, dass die Features aller drei Substanzen in einer Zusammenhangskomponente liegen, was nicht erwünscht ist.

Die Erfahrung hat gezeigt, dass es sehr schwierig ist, einen Korrelationsschwellwert zu finden, der auf der einen Seite unerwünschte Korrelationen zwischen Features verschiedener Substanzen vermeidet und auf der anderen Seite aber die erwünschten Korrelationen für Features geringer Intensität, die zu einer Substanz gehören, erhält. Es wird daher ein eher niedriger Korrelationsschwellwert (0.75-0.85) verwendet, in Verbindung mit einer Betrachtung der Kantenkonnektivität.

Da die Kantenkonnektivität in den Korrelations-Graphen von Features einer Substanz besonders hoch ist, wird ein weiterer Bearbeitungsschritt durchgeführt, der diese Eigenschaft ausnutzt und das Ziel hat, den Graphen in Komponenten besonders starker Konnektivität zu zerlegen und damit unerwünschte Korrelationen zu entfernen. Insbesondere sind dies durch einzelne Features hervorgerufene Verbindungen zwischen den stark verbundenen Subgraphen von Features verschiedener Substanzen, was in Abbildung 4.2 zu erkennen ist. Auf den Korrelations-Graphen wird daher eine Methode angewandt, um diese stark verbundenen Subgraphen zu isolieren. Als geeignet hat sich dafür der in [Hartuv und Shamir 2000] beschriebene Algorithmus zum Finden von *highly connected subgraphs* erwiesen.

Die Kantenkonnektivität (edge-connectivity) $k(G)$ eines (ungerichteten) Graphen G ist die minimale Anzahl k von Kanten, deren Entfernung in einem nicht mehr zusammenhängenden Graphen resultiert. In [Hartuv und Shamir 2000] wird definiert: Ein Graph G mit $n > 1$ Knoten wird *stark verbunden* (*highly connected*) genannt, wenn $k(G) > \frac{n}{2}$. Ein stark verbundener Subgraph (*highly connected subgraph (HCS)*) ist ein induzierter Subgraph $H \subseteq G$, so dass H stark verbunden ist. Der von den Autoren vorgestellte Algorithmus zum Finden solcher Subgraphen wird *HCS-Clustering* genannt. Nach der Definition der Autoren wäre ein Graph mit $n = 2$ Knoten, welche mit einer Kante verbunden sind, nicht stark verbunden. In der hier vorliegenden Arbeit wird dieser Fall jedoch als stark verbunden betrachtet.

Der in [Hartuv und Shamir 2000] vorgestellte Algorithmus zum HCS-Clustering beruht auf der rekursiven Anwendung eines minimalen Schnittes (minimum cut). Die Laufzeit wird mit $2 \cdot N \cdot f(n, m)$ angegeben, wobei N die Anzahl der gefundenen stark verbundenen Subgraphen ist und $f(n, m)$ die Zeitkomplexität der Berechnung eines minimalen Schnittes in einem Graphen mit n Knoten und m Kanten. Die Laufzeiten zur Berechnung eines minimalen Schnittes in einem ungewichteten Graphen werden mit $\mathcal{O}(n \cdot m)$ (deterministisch) und $\mathcal{O}(m \cdot log^3 n)$ (randomisiert) angegeben. Für diese Arbeit wurde eine Implementierung des HCS-Clusterings im R-Paket RBGL (`http://www.bioconductor.`

`org/packages/bioc/html/RBGL.html`) genutzt. Die Autoren benutzen eine optimierte Methode zur Berechnung der HCS, eine Laufzeit ist jedoch nicht angegeben.
Das Ergebnis des HCS-Clusterings für den Graph in Abbildung 4.2 rechts oben, ist in derselben Abbildung darunter dargestellt. Die stark verbundenen Subgraphen enthalten in diesem Fall genau jene Features, die zu einer Substanz gehören.
Diese Methode wird, wenn vom Benutzer gewählt, *vor* der regelbasierten Annotation durchgeführt. Diese wird dann auf den durch die Korrelationsanalyse gebildeten Feature-Gruppen durchgeführt. Die durch die Korrelationsanalyse erhaltene Gruppierung bietet gegenüber der reinen zeitlichen Gruppierung der Features den Vorteil einer durch die Korrelation verifizierten Zuordnung und außerdem, dass Überlappungen durch koeluierende Substanzen weitestgehend ausgeschlossen sein sollten.
Das Annotationsergebnis enthält dann nicht nur die Zuordnung der durch die Regeln bekannten Addukte und Fragmente, sondern auch eine weitere Spalte, in der alle durch die Korrelationsanalyse erkannten Features einer Substanz mit derselben Nummer gekennzeichnet sind. In einer geeigneten Benutzeroberfläche könnten mit diesen Informationen alle Features einer Substanz farblich markiert oder „eingefaltet“ und durch die Darstellung der berechneten Molekülmasse und der durchschnittlichen oder maximalen Intensität dieser Features ersetzt werden.

4.4 Betrachtung der Laufzeit

Betrachtet wird die Laufzeit bei Anwendung der umfangreichsten Variante der Annotation, bei der zunächst die korrelationsbasierte Gruppierung durchgeführt wird und auf den dadurch gebildeten Feature-Gruppen die regelbasierte Annotation angewandt wird. Die Anzahl der zu annotierenden Features sei mit N bezeichnet, die Anzahl Scans in der Messung mit S.
Die Berechnung der Korrelationen wird in den einzelnen Retentionszeitfenstern durchgeführt. Im günstigsten Fall ergeben sich L nicht überlappende Retentionszeitfenster, die eine mittlere Anzahl von K ($K < N$) Features enthalten, wobei die Anzahl der zu korrelierenden Intensitätswerte s sei ($s \ll S$). Für die Berechnung der paarweisen Korrelationen in einer Messung ergibt sich daher im günstigsten Fall eine Laufzeit von $\mathcal{O}(L \cdot K^2 \cdot s)$. Im ungünstigsten Fall werden alle Features einer jeweils maximalen Länge S paarweise miteinander korreliert, bei einer Laufzeit von $\mathcal{O}(N^2 \cdot S)$.
Die Laufzeit der HCS-Zerlegung wurde mit $2 \cdot H \cdot f(n, m)$ angegeben, wobei H die Anzahl der stark verbundenen Subgraphen ist und $\mathcal{O}(m \cdot \log^3 n)$ die Zeitkomplexität[12] der Funktion f zur Berechnung eines minimalen Schnittes in einem Graphen mit n Knoten und m Kanten sei. Die Anzahl der Kanten m in einem vollständig verbundenen Graphen ist $\frac{n \cdot (n-1)}{2}$. Da ein Graph mit n Knoten in maximal n Subgraphen zerlegt werden kann,

[12] $\log^p n$ wird als kürzere Schreibweise für die Potenz eines Logarithmus $(\log n)^p$ benutzt.

ergibt sich im ungünstigsten Fall ein Laufzeit von $\mathcal{O}(n^2 \log^3 n)$. Für die Betrachtung von L Graphen mit je K Knoten ergibt sich daher eine Laufzeit für die HCS-Zerlegung von $\mathcal{O}(L \cdot K^2 \log^3 K)$, wenn der günstigste Fall angenommen wird, dass jeder der L Graphen nur eine Substanz repräsentiert und damit nur einen stark zusammenhängenden Subgraphen enthält. Muss im ungünstigsten Fall ein Graph zerlegt werden, der alle N Features repräsentiert und bis zu N stark zusammenhängende Subgraphen enthält, so beträgt die Laufzeit dafür $\mathcal{O}(N^3 \log^3 N)$.
Zur Hypothesenbildung bei der regelbasierten Annotation werden alle Kombinationen der betrachteten Features mit den insgesamt J Regeln gebildet, wobei J konstant ist und $J \ll N$ gilt. Werden in jeder der L Gruppen K Features betrachtet, so beträgt die Laufzeit dafür $\mathcal{O}(L \cdot K \cdot J)$. Das Ergebnis sind im günstigsten Fall (ohne konkurrierende Hypothesen) K Annotationen für jede Gruppe. Zur Auflösung von Konflikten muss maximal jede Kombination dieser Annotationen betrachtet werden, mit einer Laufzeit von insgesamt $\mathcal{O}(L \cdot K^2)$. Im ungünstigsten Fall werden alle Features bei der Regelanwendung betrachtet, mit einer Laufzeit von $\mathcal{O}(N \cdot J)$. Wird weiterhin der ungünstigste Fall angenommen, dass die maximale Anzahl $N \cdot J$ von konkurrierenden Hypothesen betrachtet werden muss, so ergibt sich für die Konfliktauflösung eine Laufzeit von $\mathcal{O}\left((N \cdot J)^2\right)$.
Die Erkennung der Isotopomere erfolgt über ein „Suchfenster“ konstanter Größe, entsprechend der vorgegebenen maximalen Anzahl der Isotopenpeaks (z.B. 4) und Ladungszustände (z.B. 3). Da für jedes Suchfenster möglicherweise passende Features aufgesucht werden müssen (Binärsuche mit $\mathcal{O}(\log N)$) folgt für diesen Schritt eine Laufzeit von $\mathcal{O}(N \cdot \log N)$.
Die Gesamtlaufzeit für die Annotation ergibt sich daher im ungünstigsten Fall als $\mathcal{O}(N^3 \log^3 N)$ und im günstigsten Fall als $\mathcal{O}(L \cdot K^2 \log^3 K)$, was in $\mathcal{O}(N^2 \cdot \log^3 N)$ liegt.
Bei der auf Seite 77 angegebenen Laufzeit von 19 Minuten (2.4 GHz CPU) für die Annotation eines *A.thaliana* Blattextrakts mit 5166 Features entfallen ca. vier Minuten auf die regelbasierte Annotation und ca. 15 Minuten auf die korrelationsbasierte Gruppierung. Von diesen 15 Minuten werden wiederum ca. 12 Minuten zur Berechnung der paarweisen Korrelationen der EIC benötigt, die Zerlegung in stark verbundene Subgraphen nimmt rund 3 Minuten in Anspruch.

4.5 Evaluierung

Da derzeit noch kein umfangreicher Datensatz mit vollständig annotierten Features vorliegt, konnte zur Überprüfung der Leistungsfähigkeit des Annotations-Algorithmus nur die bereits in Abschnitt 3.4.1 beschriebene Mischung aus 14 bekannten Substanzen benutzt werden (Datensatz MM14). Tabelle 8.18 im Anhang (Seite 128) zeigt die Summenformeln und monoisotopischen Molekülmassen dieser Substanzen. Für diese 14 Substanzen liegt neben der bekannten Molekülmasse eine manuelle Annotation für eine Vielzahl von Featu-

res vor, die mit der durch den Algorithmus erstellten Annotation verglichen wurde. Neben der Anzahl der annotierten Features ist dabei für die Auswertung vor allem von Interesse, inwieweit die Molekülmassen der einzelnen Substanzen richtig erkannt wurden. Die Feature-Detektion wurde mit dem centWave-Algorithmus durchgeführt, unter Benutzung des Parametersatzes B (siehe Abschnitt 3.4.1).

4.5.1 Ergebnisse

Um sowohl die Ergebnisse der regelbasierten Annotation, als auch die Zuordnungen durch die Korrelationsanalyse zu überprüfen, wurde die umfangreichste Variante der Annotation genutzt. Dafür wurde zunächst die Korrelationsanalyse für alle Features durchgeführt und auf den dadurch gebildeten Feature-Gruppen die regelbasierte Annotation angewandt. Der Schwellwert für die Korrelationsanalyse wurde auf 0.75 gesetzt. Der für die Berechnung der Abstände von Isotopomeren sowie der Massendifferenzen bei Addukten und Fragmenten benötigte Toleranzwert in m/z wurde mit 0.01 gewählt, das gleitende Fenster über die Retentionszeit mit einer Breite von zwei Sekunden. Die regelbasierte Annotation wurde mit der im Anhang (8.16) dargestellten Regeltabelle durchgeführt. Diese Regeltabelle wurde ohne Kenntnis der manuellen Annotation für die 14 Substanzen angelegt.
Durch die Korrelationsanalyse mit anschließender HCS-Zerlegung wurden 733 von den insgesamt 1034 Features in 82 Feature-Gruppen (im Weiteren auch als Korrelationsgruppen bezeichnet) eingeordnet. Die anschließende Suche nach Isotopenmustern stufte 152 Features als monoisotopisch ein, diesen wurden 218 Features als Isotopomere (entsprechend dem ersten bis vierten Isotopenpeak) zugeordnet. Durch die regelbasierte Annotation der Addukte und Fragmente wurden insgesamt 176 Features annotiert, aus denen 65 Molekülmassen abgeleitet wurden.
Von den 1034 detektierten Features werden 136 durch die manuell annotierten Daten erfasst. Von diesen 136 Features wiederum beziehen sich 91 auf Addukte bzw. Fragmente und 45 auf die dazugehörigen Isotopomere. Betrachtet wird zunächst die Annotation bezogen auf die 91 Addukte und Fragmente, Tabelle 4.3 zeigt eine Zusammenfassung dieser Auswertung. Die detaillierten Ergebnisse sind in Tabelle 8.17 im Anhang (Seite 127) dargestellt, ein Beispiel daraus – die Annotation für Biochanin – zeigt Tabelle 4.4.
84 der 91 Features wurden richtig in den durch die Korrelationsanalyse gebildeten Feature-Gruppen erfasst. Von diesen 84 Addukten bzw. Fragmente wurden 44 richtig erkannt. Für 11 der 14 Substanzen wurde die richtige Molekülmasse abgeleitet. In vier Fällen, in denen nur die Quasi-Molekülionen $[M+H]^+$ und $[M+Na]^+$ annotiert werden konnten, wurden als konkurrierende Annotationen auch die in der Regelmenge enthaltenen Clusterionen $[2M+H]^+$ und $[2M+Na]^+$ sowie $[3M+H]^+$ und $[3M+Na]^+$ angegeben, mit entsprechend halber bzw. gedrittelter Molekülmasse.
In den 14 Korrelationsgruppen, welche die manuelle annotierten Features der 14 Marker-

Substanz Nr.	Bezeichnung	Anzahl Features	Davon in der Korrelati-onsgruppe	Davon richtig annotiert	Molekül-masse richtig erkannt
1	Anissäure	4	4	2	✓ *
2	Biochanin A	8	8	5	✓
3	Ferulsäure	9	9	5	✓
4	IAA-Valin	11	11	5	✓
5	Indolacetonitril	6	5	2	✓ *
6	Indolcarbaldehyd	5	5	2	✓ *
7	Kaempferol	3	3	3	✓
8	Kinetin	8	6	2	✓ *
9	p-Coumarsäure	5	3	-	-
10	Phenylalanin-d5	3	3	-	-
11	Phenylglycin	2	2	-	-
12	Phloretin	8	7	6	✓
13	Phlorizin	10	10	7	✓
14	Rutin	9	8	5	✓
Summe		91	84	44	11

Tabelle 4.3: Zusammenfassung der Annotationsergebnisse für Addukte und Fragmente auf dem MM14 Datensatz. In den mit * markierten Fällen sind auch konkurrierende Annotationen vorhanden (siehe Text).

substanzen enthalten, sind insgesamt noch 258 weitere Features enthalten. Von diesen wurden 28 annotiert und 14 Molekülmassen abgeleitet. Die einzelnen Werte sind in Tabelle 8.19 im Anhang (Seite 129) aufgeführt.

In 68 weiteren Korrelationsgruppen, über die nichts Genaueres bekannt ist, wurden insgesamt 104 Features annotiert und daraus 40 Molekülmassen abgeleitet. Tabelle 4.5 fasst diese Verteilung der annotierten Features und Molekülmassen noch einmal zusammen.

Die Annotation der Isotopomere muss in ähnlicher Weise differenziert betrachtet werden. Von der manuellen Annotation sind 45 Isotopomere erfasst, die zu den 44 erkannten Addukten bzw. Fragmenten gehören. 41 davon wurden richtig annotiert. Vier erhielten keine Annotation, wobei in zwei Fällen die Features nicht in der Korrelationsgruppe enthalten waren und in zwei weiteren Fällen die Annotation aufgrund des auf Seite 76 beschriebenen Tests abgelehnt wurde. Alle diese vier Features hatten eine sehr geringe Intensität. In der manuellen Annotation – die auf Spektren der rein gemessenen Substanzen basiert – wurden nicht für alle Ionen alle möglichen Isotopomere beschrieben. Für alle Ionen ist der

Substanz Nr.	m/z	Retentionszeit	Intensität	Manuelle Annotation	Ergebnis des Algorithmus	Abgeleitete Molekülmasse	KG
2	229.08	540.8	2428	$[C14H13O3]^+$			2
2	253.05	540.8	1968	$[M\text{-}CH_3O]^+$			2
2	270.05	540.8	5016	$[M\text{-}CH_2]^+$			2
2	285.07	540.8	644461	$[M+H]^+$	$[M+H]^+$		2
2	307.06	540.8	56425	$[M+Na]^+$	$[M+Na]^+$		2
2	591.12	541.14	8628	$[2M+Na]^+$	$[2M+Na]^+$	284.07	2
2	607.09	540.8	5465	$[2M+K]^+$	$[2M+K]^+$		2
2	891.16	540.8	583	$[3M+K]^+$	$[3M+K]^+$		2

Tabelle 4.4: Ergebnisse der Annotation für die von Biochanin A gebildeten Features. Die Spalte KG bezeichnet die Zugehörigkeit des Features zur einer Korrelationsgruppe†.

erste Isotopenpeak angegeben, für andere, die mit höherer Intensität beobachtet wurden, auch der zweite. Je nach Substanz und gemessener Intensität können aber auch mehr Isotopenpeaks beobachtet werden. So wurden neben den 45 bereits bewerteten, zusätzlich auch 35 weitere Isotopomere zugeordnet, zum Teil bis zum vierten Isotopenpeak. Die 35 weitere Annotationen der Isotopomere bezogen sich in allen Fällen auf den richtigen Ladungszustand und wurden daher als zutreffend bewertet. Den erwähnten 28 Features (siehe Tabelle 4.5), die außerdem in dem Korrelationsgruppen 1-14 enthalten sind, wurden 45 Isotopomere zugeordnet, den 104 Features in den Korrelationsgruppen 15-82 nochmals 84.
Bei der als erstes durchgeführten Korrelationsanalyse wurden zunächst 69 Zusammenhangskomponenten gebildet. In 18 Fällen waren diese Zusammenhangskomponenten schon stark zusammenhängend. In 26 Fällen wurde nur einzelne Features von einer ansonsten starken Zusammenhangskomponente abgetrennt, davon waren vier Features betroffen, die eigentlich zur entsprechenden Substanz gehören. In 25 Fällen wurde auch mehr als ein Feature abgetrennt, welche dann zum Teil kleinere Zusammenhangskomponenten ergaben. Das Resultat waren die erwähnten 82 Korrelationsgruppen.

4.5.2 Diskussion

Das wichtigste Ergebnis ist zunächst die richtige Erkennung von elf Molekülmassen. Für p-Coumarsäure waren zwar drei Features vorhanden, die zur richtigen Annotation geführt hätten, jedoch traten diese, insbesondere das $[M+H]^+$ Ion, nur mit sehr geringer Intensität auf. Die Intensitätskorrelationen dieser Features lagen unter dem Schwellwert und

†Die laufende Nummerierung der Korrelationsgruppen ist in der Programmausgabe anders, für eine übersichtliche Darstellung wurde diese an die Nummerierung der Substanzen angepasst.

Betrachtete Korrelations-gruppen†	Beschreibung	Anzahl annotierter Features	Anzahl abgeleiteter Molekül-massen
1-14	Annotierte Features, die in der manuellen Annotation enthalten sind	44	11
	Annotierte Features, die nicht in der manuellen Annotation enthalten sind	28	14
15-82	Annotierte Features	104	40
1-82	Summe	176	65

Tabelle 4.5: Verteilung der 176 insgesamt annotierten Features

dementsprechend wurden diese Features nicht der Korrelationsgruppe der Substanz zugeordnet. Diese enthielt somit nur zwei Fragmentionen und ein Clusterion, was mit den gegebenen Regeln nicht für die Annotation ausreichte.

Sowohl für Phenylglycin als auch für das deuterierte Phenylalanin konnten nicht die für die Annotation benötigten, mindestens zwei durch die Regeln erklärten Ionen, beobachtet werden. Auch hier enthielten die entsprechenden Korrelationsgruppen nur Fragmentionen bzw. nur ein Quasi-Molekülion, infolgedessen konnte für diese beiden Substanzen keine Annotation erstellt werden.

In sieben Fällen lieferte die regelbasierte Annotation ein eindeutiges Ergebnis ohne weitere falsch positive Annotationen. Die in vier Fällen auftretenden, konkurrierenden Annotationen sind durch das bereits erwähnte Problem der identischen m/z-Differenz von Quasi-Molekülionen und entsprechenden Clusterionen gegeben. Dieses Problem tritt nur dann auf, wenn nur wenige Quasi-Molekülionen beobachtet werden, für die gleichzeitig auch Regeln mit den gleichen Kationen für die Bildung von Clusterionen vorgegeben wurden. Zu diesem Zeitpunkt steht jedoch noch kein umfangreicher Datensatz zur Verfügung, mit dem überprüft werden kann, ob die Implementierung einer „harten“ Regel für dieses Problem nicht auch in Einzelfällen zu fehlerhaften Annotationen führt. Die Auflösung dieses Konflikts wird daher noch der Interpretation durch den Benutzer überlassen. Sobald Quasi-Molekülionen zusammen mit mindestens einem weiteren Ion eines anderen Typs, d.h. ein Clusterion oder ein Fragment, beobachtet werden, greift die Kardinalitätsregel und es tritt (zumindest in den beobachteten Fällen) keine Mehrdeutigkeit auf. Für die Annotation von Anissäure beispielsweise hätte die Erweiterung der Regelmenge um das Ion $[M+2Na-H]^+$ die Eindeutigkeit der Annotation bewirkt. In zwei Fällen (Ferulsäure und Phlorizin) weicht die erkannte Molekülmasse um m/z 0.01 vom Sollwert ab. Um die

Frage zu klären, es sich hierbei nur um zufällige Abweichungen handelt, oder ob andere Modelle als die bislang genutzte Mittelwertbildung aus den Molekülmassen der erkannten Ionen zu konsistent besseren Ergebnissen führen, ist eine umfangreichere Messung von bekannten Substanzen nötig. Eine weitere Möglichkeit wäre beispielsweise, nur die ein oder zwei Features mit der höchsten Intensität für diese Berechnung zu nutzen.
Die Zuordnung der Isotopomere war für alle in den Korrelationsgruppen enhaltenen und mittels der manuelle Annotation bewerteten Fällen richtig. Mehrdeutigkeiten oder falsch positive Annotation traten dabei nicht auf. Über die Annotationen für die nicht in der manuellen Annotation enthaltenen Features lässt sich, wie auch für die dort annotierten Addukte und Fragmente, keine Aussage treffen.
Die Korrelationsanalyse lieferte auf dem Datensatz eine richtige Gruppierung bezüglich der manuell annotierten Features der einzelnen Substanzen. Problematisch hierbei sind nur Features mit sehr niedriger Intensität, die dann aufgrund geringer Korrelation nicht zugeordnet werden können. Dieser Fall trat bei sieben Features auf, wobei drei davon unter dem Korrelationsschwellwert lagen und vier bei der HCS-Zerlegung abgetrennt wurden, da sie nur einfach mit der Zusammenhangskomponente verbunden waren. Alle anderen Zerlegungen können nicht bewertet werden, da die entsprechenden Features unbekannt sind.

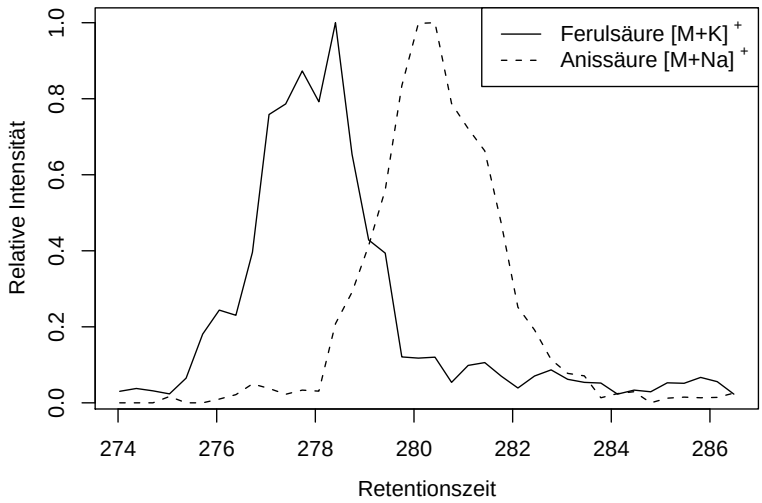

Abbildung 4.3: Extrahierte Ionen Chromatogramme für je ein Ion von Anissäure und Ferulsäure aus dem Datensatz MM14.

Die Funktion der HCS-Zerlegung für koeluierende Substanzen kann mit dem Datensatz MM14 nicht bewertet werden, da solche Fälle nicht auftraten. Anissäure und Ferulsäure eluieren zwar zu einer recht ähnlichen Retentionszeit ($\approx$ 280 bzw. 278 Sekunden, siehe Abbildung 4.3), dieser zeitliche Abstand reichte jedoch aus, um die entsprechenden Features

schon bei der zeitlichen Gruppierung in verschiedene Gruppen einzuordnen. Die chromatographischen Peaks überlappen zudem nur geringfügig, so dass auch bei einem größer gewählten Fenster für die zeitliche Gruppierung die Features schon vor der HCS-Zerlegung in verschiedenen Zusammenhangskomponenten liegen. Der Korrelationskoeffizient für die chromatographischen Intensitäten beträgt im gezeigten Beispiel -0.07.

4.5.3 Laufzeit

Die Gesamtlaufzeit der Annotation auf dem Datensatz MM14 mit 1034 Features wurde auf einem 2.4 GHz Prozessor mit rund einer Minute gemessen, wovon ca. je 30 Sekunden auf die korrelationsbasierte Gruppierung bzw. die regelbasierte Annotation entfallen.

4.6 Zusammenfassung

Die Annotation zusammengehöriger Features liefert Gruppierungen der detektierten Features entsprechend der zugehörigen Verbindungen und erleichtert so die weitere Auswertung. Aus den zu einer Verbindung gehörenden Features kann mithilfe der vorgegebenen Ionisierungsregeln die entsprechende Molekülmasse abgeleitet werden. Diese Informationen können für eine kompakte Repräsentation des Datensatzes genutzt werden und erleichtern weitere Schritte der Interpretation wie die Massendekomposition [Böcker und Lipták 2005, Grange und Sovocool 2008] oder die Molekülsuche in Datenbanken [Horai et al. 2008, Iijima et al. 2008]. Der vorgestellte Algorithmus unter Verwendung der regelbasierten Annotation und der Korrelationsanalyse stellt die erste frei verfügbare Lösung in dieser Richtung dar.
Eine Evaluierung des Verfahrens wurde auf einer LC/MS Messung von 14 bekannten Verbindungen durchgeführt. Die mithilfe des korrelationsbasierten Verfahrens durchgeführte Feature-Gruppierung ergab für die Features aus dem Referenzdatensatz eine in den meisten Fällen (84 von 91) richtige Gruppierung. Problematisch sind hier Features mit besonders niedriger Intensität, die aufgrund der nur schwachen Korrelation nicht in die Feature-Gruppe aufgenommen werden.
Die regelbasierte Erkennung der Molekülmasse ergab in 11 von 14 Fällen ein richtiges Ergebnis, wobei in vier Fällen auch konkurrierende Annotationen vorhanden waren. Ursache für die nicht erkannten Molekülmassen war in zwei Fällen, dass keine bzw. ungenügend viele der vorgegebenen Regeln anwendbar waren und in einem Fall eine fehlende Featurezuordnung aufgrund des erwähnten Problems bei sehr niedriger Featureintensität.
Die Leistungsfähigkeit der regelbasierten Annotation hängt entscheidend von der zur Verfügung stehenden Regelmenge ab. Hier sind Verbesserungen unter Verwendung einer umfangreicheren Regelmenge denkbar, welche z.B. mit Hilfe einer Datenbank von annotierten Spektren erlernt werden könnte. Weiterhin sollte der Einsatz von Kriterien zum

Ausschluss von konkurrierenden Annotationen geprüft werden. Zur Weiterentwicklung des Algorithmus ist in jedem Fall eine größere Basis an annotierten Features notwendig, welche derzeit leider noch nicht zur Verfügung steht.

5 Evaluierung von Alignment-Methoden

Für viele Anwendungen im Metabolomik-Bereich ist es erforderlich, die Intensitäten von gemessenen Substanzen über mehrere, manchmal hunderte, LC/MS-Messungen hinweg zu vergleichen. Typische Experimente solcher Art im Bereich der Pflanzen-Metabolomik sind z.B. die Untersuchung von *Arabidopsis thaliana* Stoffwechsel-Mutanten [Böttcher et al. 2008], diurnaler Zyklen [Gibon et al. 2006], Lichtüberempfindlichkeit [Kolotilin et al. 2007] oder Wundstress [Boccard et al. 2007]. Um die Vergleichbarkeit der Messungen zu ermöglichen ist es nötig, die auftretenden zeitlichen Abweichungen bei den chromatographischen Trennungen und auch der m/z-Werte zu berücksichtigen bzw. gegebenenfalls zu kompensieren. Während die Abweichungen in m/z normalerweise sehr gering sind, kommt es bei den chromatographischen Trennungen, bedingt durch das Altern der Säule, Temperatur- und Druckschwankungen, zu zum Teil erheblichen Schwankungen in der Retentionszeit der eluierenden Substanzen. Insbesondere bei Substanzen, die ähnliche Retentionszeiten aufweisen, kann es dabei auch vorkommen, dass sich die Reihenfolge verändert, in der die Substanzen eluieren. Es existieren verschiedene Ansätze, um diese Abweichungen in der Retentionszeit zu korrigieren (*Retentionszeitkorrektur*) und danach eine Gruppierung der Features durchzuführen. Andere Verfahren führen keine systematische Korrektur durch, sondern versuchen, korrespondierende Features unter Berücksichtigung von tolerierten Abweichungen direkt zu gruppieren. Der gesamte Verarbeitungsprozeß, der schließlich gruppierte Features über alle Messungen liefert, wird *Alignment* genannt. Werden mehr als zwei Messungen gruppiert, wird dies als *Multiples Alignment* bezeichnet.

Die Zuverlässigkeit des Alignments ist von besonderer Bedeutung. Falsch alignierte Features können die für das Ergebnis des Experiments entscheidende statistische Auswertung der Feature-Intensitäten beeinträchtigen, bzw. sogar verfälschen. Gleichzeitig ist es nahezu unmöglich, das Alignment-Ergebnis manuell zu verifizieren, da dazu die Intensität jedes einzelnen von oft tausenden Features in jedem Experiment (manchmal hundert und mehr) überprüft werden müsste. Aus diesen Gründen ist sowohl eine geeignete Wahl des Alignment-Algorithmus zu treffen, als auch eine sorgfältige Parameteroptimierung durchzuführen.

Mittlerweile ist eine Vielzahl von Alignment-Algorithmen verfügbar (ein Überblick ist z.B. in [Katajamaa und Oresic 2007] gegeben), über deren Leistungsfähigkeit mangels eines objektiven Evaluierungskriteriums wenig bekannt ist. Um einen Vergleich der Algorithmen zu ermöglichen, wurden daher umfangreiche LC/MS-Referenzdatensätze erstellt und eine Evaluierung von Algorithmen mit diesen Datensätzen durchgeführt. Evaluiert wurden nur solche Programme, die ein multiples Alignment beherrschen und deren Implementierung frei verfügbar ist. Dies sind derzeit MZmine, XCMS, XAlign und der MapAlignment-Algorithmus von OpenMS.

Im Folgenden werden die Erstellung von Referenzdaten und die Evaluierung dieser vier Algorithmen auf den Metabolomik-Datensätzen diskutiert, was den Eigenanteil an den in [Lange et al. 2008] veröffentlichten Ergebnissen beinhaltet. Weiterhin werden die gemeinsam mit den anderen Autoren von [Lange et al. 2008] entwickelten Vergleichskriterien erläutert. Die verwendeten Referenzdatensätze wie auch die R-Skripte zur Evaluierung sind unter `http://msbi.ipb-halle.de/msbi/caap` verfügbar.

5.1 Definitionen und Begriffe

Das Alignment-Problem wird häufig in zwei Teilproblemen betrachtet:

1. dem Finden einer geeigneten Transformation (meist als Retentionszeitkorrektur bezeichnet) für die Retentionszeit von LC/MS-Messungen, so dass korrespondierende Features nach Anwendung der Transformation eine ähnliche Retentionszeit besitzen, sowie

2. dem eigentliche Gruppieren der korrespondierenden Features über alle LC/MS-Messungen hinweg.

Eine solche Gruppe von Features, welche die Koordinaten und Intensitäten von korrespondierenden Features (dasselbe Ion derselben Substanz) für alle alignierten Messungen beschreibt, wird *Konsensus-Feature* genannt. Die Menge aller Konsensus-Features bildet die *Konsensus-Liste*, in der somit Informationen über den Zusammenhang aller detektierten Features für die betrachteten LC/MS-Messungen enthalten sind.
Es kommt vor, dass Konsensus-Features nicht für jede der alignierten LC/MS-Messungen ein entsprechendes Feature enthalten. Dies ist dann der Fall, wenn das entsprechende Feature nicht in allen Messungen detektiert wurde (da nicht vorhanden bzw. Intensität zu gering) oder aber die Features beim Alignment nicht in allen Fällen richtig zugeordnet wurden. XCMS und MZmine enthalten Methoden, mit denen eine „Nachsuche“ auf den Rohdaten vorgenommen werden kann, um derartige Lücken zu füllen. Von dieser Möglichkeit wurde jedoch bei der durchgeführten Evaluierung kein Gebrauch gemacht.
Weiterhin können je nach Alignment-Strategie Fälle auftreten, in denen die zunächst gebildeten Konsensus-Features mehr Features enthalten, als LC/MS-Messungen aligniert wurden. Dies tritt z.B. dann auf, wenn sehr eng beieinander liegende Features aus einer Messung gemeinsam (statt nur eines von beiden) einem Konsensus-Feature zugeordnet werden. Dieser Effekt wird z.B. gelegentlich für Isomere einer Substanz beobachtet, bei denen die gebildeten Features exakt die gleiche Masse besitzen und sich nur durch eine geringfügig andere Retentionszeit unterscheiden. Konflikte solcher Art werden im Allgemeinen dadurch aufgelöst, indem aus mehreren in Frage kommenden Features dasjenige Feature ausgewählt wird, dessen Koordinaten näher am Mittelwert des Konsensus-Features

liegen. XCMS bietet anstelle dessen auch die Option, die Auswahl zugunsten des Features mit der höheren Intensität durchzuführen. Diese Option wurde bei der Evaluierung nicht benutzt.
Außer den beschriebenen, „natürlich“ vorkommenden Problemen bei der Zuordnung einzelner Features können außerdem die verschiedensten Arten von Fehlern beim Alignment auftreten. Eine Ursache dafür kann z.B. eine ungenügende Retentionszeitkorrektur sein, andererseits aber auch eine „Überkorrektur“ bzw. Verzerrung der Retentionszeit. In beiden Fällen erhöht sich die Wahrscheinlichkeit der Fehlzuordung von Features. Hierbei wird zwischen falsch positiven (fehlerhaft alignierten) Features und falsch negativen (fehlerhaft nicht alignierten) Features unterschieden.
Im Ergebnis des Algorithmus – der Konsensus-Liste – sollten zusammengehörige Features in einem Konsensus-Feature gruppiert sein, ein Aufsplitten in mehrere Konsensus-Features soll vermieden werden, ebenso wie die Gruppierung von nicht-zusammengehörigen Features. Im Abschnitt 5.4 werden zwei Maße eingeführt, um die Qualität einer bestimmten Konsensus-Liste in Bezug zu einer als optimal bewerteten Konsensus-Liste, dem *Referenzdatensatz*, einzuschätzen.

5.2 Alignment Ansätze

Der Vollständigkeit halber sei erwähnt, dass es neben dem häufig verwendeten Verfahren, zunächst die Feature-Detektion auf den LC/MS-Messungen durchzuführen und anschließend die detektierten Features zu alignieren, auch die Möglichkeit gibt, direkt auf den Rohdaten zu arbeiten, diese zu alignieren und dann z.B. Unterschiede zwischen den gemessenen Datensätzen zu finden. Verschiedene Ansätze in dieser Art wurden bereits beschrieben, siehe [Prakash et al. 2006, Bylund et al. 2002, Prince und Marcotte 2006, Baran et al. 2006, Listgarten und Emili 2005, Listgarten et al. 2007]. Sie besitzen den Vorteil, dass Fehler bei der Feature-Detektion umgangen werden können. Ein Nachteil ist jedoch, dass die meisten dieser Algorithmen nur für das Alignment und den Vergleich von genau zwei Messungen ausgelegt sind und daher nicht ohne weiteres für ein multiples Alignment verwendet werden können.
Zur Durchführung des Alignments auf LC/MS-Feature-Ebene wurden verschiedene Verfahren vorgestellt [Radulovic et al. 2004, Katajamaa et al. 2005, Li et al. 2005, Zhang et al. 2005, Jaitly et al. 2006, Bellew et al. 2006, Smith et al. 2006, Wang et al. 2007] von denen einige als eigenständige Methode entwickelt worden und andere in größere Frameworks zur Analyse massenspektrometrischer Daten eingebettet sind. Für den in [Lange et al. 2008] durchgeführten Vergleich wurden die Alignment-Algorithmen innerhalb der Frameworks msInspect [Bellew et al. 2006], MZmine [Katajamaa et al. 2005], OpenMS [Lange et al. 2007] und XCMS [Smith et al. 2006] betrachtet, sowie die eigenständigen Programme SpecArray [Li et al. 2005] und XAlign [Zhang et al. 2005].

Den Alignment-Algorithmus von MZmine ausgenommen, führen alle Programme eine Retentionszeitkorrektur durch, um die durch die Chromatographie entstandenen zeitlichen Abweichungen zu korrigieren. Alle zuletzt aufgeführten Algorithmen erfüllen die Bedingung ein multiples Alignment durchführen zu können und sind als Implementierung frei verfügbar. Die für Proteomik entwickelten Programme SpecArray und msInspect stellten sich als (zumindest in den derzeitigen Versionen 2.1 bzw. 1.0.1) für Metabolomik Daten nicht geeignet heraus [Lange et al. 2008] und werden daher hier nicht weiter betrachtet. Der Grund für diese Inkompatibilität mit Metabolomik-Daten ist dabei wahrscheinlich am ehesten in der gezielten Entwicklung der Programme für Proteomik-Daten zu suchen, denn technisch sind sich die erzeugten Datensätze ähnlich. Die chromatographische Trennmethode ist vergleichbar, es wird jeweils C_{18}-Umkehrphasen Chromatographie verwendet. Der Hauptunterschied dürfte in den gemessenen Massenbereichen liegen. Im Metabolomik-Bereich ist die Messung von m/z 50–1000 üblich, für Proteomik-Experimente eher Bereiche von m/z 400–2000. Weiterhin werden bei der Messung von Metaboliten vorwiegend einfach geladene Ionen beobachtet, während bei Proteinen bzw. Peptiden auch deutlich höhere Ladungszustände gemessen werden [Wolters et al. 2001, Chen et al. 2006]. Durch die größeren Massen und höheren Ladungszustände ergeben sich komplexere Isotopenmuster, die Identifizierung des C_{12}-Peaks eines Ions –und damit die Feststellung der exakten monoisotopischen Masse– ist im Gegensatz zu Metabolomik-Daten teilweise problematisch [Baldwin 2004]. In Folge dessen unterscheiden sich die beim Alignment zu tolerierenden m/z-Abweichungen um eine Größenordnung. In [Lange et al. 2008] wurden zum Alignment der Metabolomik-Daten m/z-Toleranzen von 0.01 bis 0.1 verwendet, für die Proteomik-Datensätze jedoch 1.5 bis 2.5 m/z. Die für Metabolomik konzipierten Algorithmen können sich daher beim Alignment weitaus mehr an den m/z-Koordinaten der Features orientieren.

Da der Vergleich der Programme ausschließlich bezüglich des Alignments erfolgen sollte, mussten einige der Programme (MZmine und XCMS) so modifiziert werden, dass die vorbereiteten Feature-Listen eingelesen werden, anstatt selbst die Feature-Detektion vorzunehmen. Dadurch wurde eine zusätzliche Variabilität vermieden, die den Vergleich erschwert hätte. Im Folgenden wird das grundlegende Funktionsprinzip der vier evaluierten Alignment-Algorithmen (MZmine, OpenMS, XAlign und XCMS) erläutert.

5.2.1 MZmine

Allgemeine Informationen zu MZmine [Katajamaa und Oresic 2005] (Version 0.60) finden sich in Abschnitt 3.2.3. Für die Evaluierung der Alignments wurde MZmine modifiziert, um den für das Alignment ansonsten obligatorischen Feature-Detektions-Schritt zu überspringen und stattdessen die vorbereiteten Feature-Listen zu importieren.

MZmine benutzt für das Alignment keine Retentionszeitkorrektur, sondern verwendet nur

einen relativ einfachen Algorithmus zur Gruppierung. Verwaltet wird dazu eine anfänglich leere *master feature list*, die nach und nach mit den Features aus den Messungen gefüllt wird. Eine Distanz-Funktion wird benutzt, um die Ähnlichkeit zwischen einem Feature und den Einträgen in der master-list zu berechnen. Als Distanz wird die gewichtete Summe aus den Distanzen in *m/z* und der Retentionszeit (RT) verwendet, wobei RT mit 1 gewichtet wird und *m/z* mit 10. Wenn die Distanz zwischen dem Feature und dem am besten passendsten Eintrag aus der master-list ausreichend gering ist, wird das Feature diesem Eintrag zugeordnet und ansonsten als neuer Eintrag an die master-list angehängt. Ausreichend gering heißt, das die Distanz kleiner sein muss als eine vorgegebene Maximaldistanz, die sich wiederum aus der vom Benutzer eingegebenen maximal tolerierten Abweichung in *m/z* und RT ergibt.

5.2.2 MapAlignment (OpenMS)

OpenMS [Sturm et al. 2008](Version 1.0) ist ein Open-Source-Framework zur Analyse von Massenspektrometriedaten. Enthalten ist ein für das Alignment von Proteomik-Daten konzipierter Algorithmus (*MapAlignment*, [Lange et al. 2007]), der sich im durchgeführten Vergleich aber auch als bedingt für Metabolomik-Daten geeignet erwiesen hat.
Die Feature-Listen aus den einzelnen Messungen werden dabei auf sternförmige Art paarweise aligniert, wobei die umfangreichste Feature-Liste das Zentrum bildet. Zur Korrektur der Abweichungen in *m/z* und RT versucht der Algorithmus zunächst eine optimale affine Abbildung zu bestimmen, die möglichst viele Features aus einer Liste so nahe wie möglich auf Features einer anderen Liste abbildet. Dieser Ansatz wird als *pose-clustering* bezeichnet. Nachdem so eine erste Schätzung der Abbildung erfolgt ist und auf eine der Listen angewandt wurde, werden die dann übereinstimmenden Features als „landmarks“ betrachtet und für die Berechnung einer linearen Regression benutzt. Im abschließenden Konsensus-Schritt des Algorithmus wird, basierend auf Nächster-Nachbar-Suche, die endgültige Konsensus-Liste berechnet, wobei die abweichungskorrigierten Feature-Listen benutzt werden. Wichtige Eingabeparameter sind *mz_bucket_size*, welcher die Größe der „buckets“ bei der Suche nach der affinen Transformation bestimmt, sowie die Parameter *precision_RT* und *precision_MZ*, die die maximal erlaubte Distanz korrespondierender Features für den Konsensus-Schritt vorgeben.

5.2.3 XAlign

XAlign wurde für das Alignment von Proteomik LC/MS Daten entwickelt, angegeben ist jedoch auch die Anwendbarkeit auf Metabolomik Daten [Zhang et al. 2005]. XAlign ist auf direkte Anfrage beim Autor als Windows-Programm ohne Quelltext erhältlich. Die verwendete Programmversion ist mit dem Datum 03.09.2007 gekennzeichnet.

XAlign berechnet in einem ersten Schritt zunächst ein „grobes“ (*„gross“*) Alignment, um systematische Abweichungen in RT zu korrigieren. In einem zweiten Schritt, *Micro-Alignment* genannt, wird die endgültige Konsensus-Liste bestimmt. Für das „grobe“ Alignment werden über alle Feature-Listen hinweg *signifikante* Features gesucht. Ein *signifikantes* Feature bezeichnet ein Feature, welches in allen Feature-Listen gefunden wurde und außerdem für einen bestimmten Bereich das Feature mit der höchsten Intensität darstellt. Die Größe dieses Bereichs ist abhängig von der vom Benutzer eingegeben zulässigen Toleranz in *m/z* und RT. Die Feature-Liste, deren signifikante Features die geringste zeitlichen Abweichung zu den gemittelten RT-Werten der signifikanten Features zeigen, wird als Referenz-Liste gewählt. Unter Benutzung der signifikanten Features werden die zeitlichen Abweichungen für jede Feature-Liste gegen die Referenz-Liste nach der Berechnung einer linearen Regression korrigiert. In der Micro-Alignment-Phase werden sukzessive die Features aus allen Listen gruppiert und bilden schließlich die Konsensus-Liste.

5.2.4 XCMS

Der in [Smith et al. 2006] vorgestellte Alignment-Algorithmus für Metabolomik Daten ist in dem bereits in Abschnitt 3.2.3 beschriebenen R-Paket *XCMS*(Version 1.12.1) enthalten. In einem Gruppierungsschritt werden dabei zunächst Features aus allen Listen mit ähnlichem *m/z*-Wert betrachtet, indem Intervalle fest vorgegebener Breite (Parameter *mzwid*) gebildet werden. Für alle Features in solch einem *m/z*-Intervall wird eine Schätzung der lokalen Feature-Dichte über der Zeitachse mittels Kernel Density Estimation durchgeführt. Ausgehend von den lokalen Maxima der Dichteschätzung werden dann die Features, deren Retentionszeit im Bereich bis zum nächsten lokalen Minimum liegt, jeweils einem Konsensus-Feature zugeordnet. Nach diesem Gruppierungsschritt existiert damit schon eine Konsensus-Liste. Abhängig davon, ob stärkere (systematische) zeitliche Abweichungen zu erwarten sind, kann anschließend ein Retentionszeitkorrektur-Schritt durchgeführt werden. Dafür werden aus den schon gruppierten Features solche ausgewählt („landmarks“), die bestimmten Anforderungen genügen: diese Features sollen in mindestens einer bestimmten Anzahl von Feature-Listen vorhanden sein, aber auch höchstens in allen Feature-Listen genau je einmal. Für diese Konsensus-Features wird nun die zeitliche Abweichung der einzelnen Features gegenüber dem jeweiligen zeitlichen Mittelwert des Konsensus-Features betrachtet. Es kann eine lineare oder nichtlineare (local polynomial regression fitting, R-Funktion *loess*) Methode gewählt werden, um diese Abweichungen zu beschreiben und zu korrigieren.
Nach dieser Korrektur wird wieder der Gruppierungsschritt angewendet, um eine Neuzuordung der Features, welche nunmehr andere Retentionszeiten haben, zu ermöglichen. Bei sehr starken Retentionszeitabweichungen kann das gesamte Verfahren mehrfach durchgeführt werden, mit der Hoffnung, in jedem Schritt mehr Features alignieren zu können.

Wichtige Parameter sind, neben dem erwähnten *mzwid* als Intervallbreite, der Parameter *bw*, der die Bandbreite für die Kernel Density Estimation vorgibt und damit indirekt die tolerierte zeitliche Abweichung beim Gruppierungsschritt, sowie der Glättungsparameter *span* für die nichtlineare Schätzung der Retentionszeitabweichung.
Bei den Versuchen wurde festgestellt, dass es schwer vorherzusagen ist, ob die Retentionszeitkorrektur nötig ist, bzw. ob sie das Ergebnis verbessert. Für die Evaluierung sind daher die Ergebnisse sowohl mit als auch ohne Retentionszeitkorrektur angegeben.

5.3 Erstellung des Referenzdatensatzes

Als Grundlage für die Erstellung des Referenzdatensatzes wurde ein typisches Metabolomik Experiment ausgewählt. Blattextrakte von *Arabidopsis thaliana* wurden auf zwei verschiedenen LC/ESI-QTOF-MS Gerätekonfigurationen gemessen.

5.3.1 LC/MS Messungen

Datensatz M1:
Gemessen wurden elf verschiedene Blattextrakte vom *A. thaliana* Wildtyp und verschiedenen Mutanten, mit je vier technischen Replikaten. Die chromatographische Separation wurde auf einem Kapillar LC-System von Dionex durchgeführt. Die eluierenden Substanzen wurden von *m/z* 75 bis 1000 mittels eines API QSTAR Pulsar i (Applied Biosystems/MDS Sciex) Massenspektrometers detektiert, betrieben mit einer „Ionspray“ Elektrospray-Ionenquelle im positiven Ionisierungsmodus. Die Akkumulationszeit eines Scans betrug 2 Sekunden. Die Massenauflösung für ein Kalibrierungs-Peptid (bei *m/z* 829) betrug 8500.
Datensatz M2:
Gemessen wurden sechs verschiedene Blattextrakte vom *A. thaliana* Wildtyp und verschiedenen Mutanten, mit je vier technischen Replikaten. Die chromatographische Separation wurde auf einem HPLC-System von Agilent durchgeführt. Die eluierenden Substanzen wurden von *m/z* 100 bis 1000 mittels eines MicrOTOF-Q (Bruker Daltonics) Massenspektrometers detektiert, betrieben mit einer „Apollo II“ Elektrospray-Ionenquelle im positiven Ionisierungsmodus. Die Akkumulationszeit eines Scans betrug 1,5 Sekunden. Die Massenauflösung für ein Kalibrierungs-Peptid (bei *m/z* 829) betrug 14000.
Alle Daten wurden im Centroid-Modus mittels der jeweiligen Konvertierungs-Software exportiert. Die Feature-Detektion wurde mit dem in Kapitel 3 beschriebenen centWave-Algorithmus durchgeführt. Für den Datensatz M1 wurden dafür die Parameter *peakwidth*=(20,50), $snthresh = 5$, $ppm = 120$ genutzt. Zur Feature-Detektion auf dem Datensatz M2 wurden die gleichen Parameter benutzt, mit Ausnahme einer geringer eingestellten Massenabweichung von $ppm = 30$.

5.3.2 Alignment mithilfe der korrelationsbasierten Gruppierung

Für Metabolomik LC/MS-Daten sind –im Gegensatz zur der in [Lange et al. 2008] verwendeten SEQUEST-Identifikation für Proteomik-Daten– noch keine Datenbanken vorhanden, die eine umfangreiche Annotation der enthaltenen Substanzen ermöglichen und somit zur Erstellung eines Referenzdatensatz benutzt werden könnten.

Um in Ermangelung einer vollständigen chemischen Identifizierung trotzdem eine Evaluation der Alignment-Methoden auf Metabolomik-Daten zu ermöglichen, wurde mit den vorhandenen Mitteln eine relative Annotation von Features durchgeführt – mit dem Ziel, Gruppen von korrespondierenden Features zu finden, auch ohne deren genaue Identität zu kennen. Die Idee hierbei war es, mithilfe der in Abschnitt 4.3.2 beschriebenen Korrelationsanalyse zunächst unabhängig in jeder Messung Features einer Substanz zu gruppieren und in einem weiteren Schritt über alle Messungen hinweg die jeweils korrespondierenden Feature-Gruppen zu verbinden. Der Referenzdatensatz wird damit ebenfalls über eine Art Alignment gebildet, welches jedoch nicht wie üblich auf der Zuordnung individueller Features, sondern ganzer Feature-Gruppen beruht, die über die Korrelationsanalyse gebildet wurden. Durch die Gruppierung der Features in einen chemischen Zusammenhang kann die Zuordnungssicherheit gegenüber einem „normalen“ Alignment, welches ohne dieses Vorwissen funktionieren muss, deutlich erhöht werden. Ein Nachteil ist jedoch, dass dabei nur eine Teilmenge von Features erfasst wird, nämlich solche, die in den Feature-Gruppen erfasst wurden und bei der nachfolgend geschilderten Prozedur eindeutig zuzuordnen waren. Der Schwerpunkt beim Aufbau dieses Referenzdatensatzes lag damit nicht bei der Erstellung eines möglichst vollständigen Alignments über einen „Umweg“, sondern bei der möglichst sicheren Zuordnung einer Teilmenge von Features jeder Messung zu Konsensus-Features, deren Reproduktion durch die verschiedenen Alignment-Methoden dann geprüft werden kann.

Es wurde zunächst die in Abschnitt 4.3.2 beschriebene Korrelationsanalyse mit anschließender Zerlegung in Subgraphen verwendet, um für jede LC/MS-Messung solche Features zu gruppieren, die mit sehr hoher Wahrscheinlichkeit zu einer Substanz gehören. Der Korrelations-Schwellwert wurde dafür mit 0.9 vergleichsweise hoch angesetzt. Im nächsten Schritt wurde versucht, diese Gruppen von Features über die Messungen hinweg zuzuordnen. Dazu wurde geprüft, welche Gruppen innerhalb bestimmter tolerierter Abweichungen (Datensatz M1: $\Delta\mathrm{RT} = 90$ s, $\Delta m/z = 0.02$, Datensatz M2: $\Delta\mathrm{RT} = 20$ s, $\Delta m/z = 0.01$) eindeutig Gruppen aus einer anderen Messung zugeordnet werden können. Zur Bestimmung der zeitlichen Toleranzwerte wurden die Zeitabweichungen der Datensätze über Betrachtungen der TIC's und der Abweichungen bei den eingespikten Markersubstanzen abgeschätzt. Die maximale Abweichung in m/z wurde unter Kenntnis der Genauigkeit der beiden verwendeten Massenspektrometer festgelegt.

In den verwendeten Blattextrakten von verschiedenen *Arabidopsis thaliana* Mutanten

wurden Unterschiede in maximal 5% der Metabolite erwartet (Gespräch mit C. Böttcher). Die Gruppierung wurde daher in sternförmiger Art ausgeführt, wobei jeweils eine Messung des *Arabidopsis thaliana* Wildtyps das Zentrum bildete. Ausgehend von dieser Messung wurden die Feature-Gruppen aus den anderen Messungen zugeordnet, wobei die erwähnten Toleranzwerte zum Finden der Gruppen verwendet wurde. Als Kriterium für die erfolgreiche Zuordnung wurden mindestens zwei gemeinsame Features zwischen den verglichenen Gruppen gefordert. Generell wurde nur solche Gruppenzuordnungen weiterverwendet, bei denen der Vergleich eindeutig war, d.h. keine anderen Features innerhalb der Toleranzwerte in Frage kamen.
In einem nächsten Schritt wurden die im sternförmigen Vergleich gefundenen Gruppenzuordnungen auf ihre Gemeinsamkeiten überprüft und nur die Schnittmenge weiterverwendet. Beispielsweise kann die Zuordnung der Gruppen zwischen dem Zentrum des Alignments (0) und der Messung 1 fünf gemeinsame Features in den Gruppen ergeben, aber dieselbe Gruppe aus dem Zentrum beim Vergleich mit der Messung 2 nur drei gemeinsame Gruppenfeatures haben.
Alle erhaltenen Gruppenzuordnungen, die sich über mindestens vier Messungen erstrecken, wurden zur Bildung des Referenzdatensatzes benutzt. Jede der erhaltenen Gruppen lieferte somit zumindest ein, häufig jedoch mehrere Konsensus-Features. Abbildung 5.1 zeigt die Verteilung der Länge dieser Konsensus-Features. Ein großer Anteil erstreckt sich dabei über alle Messungen.

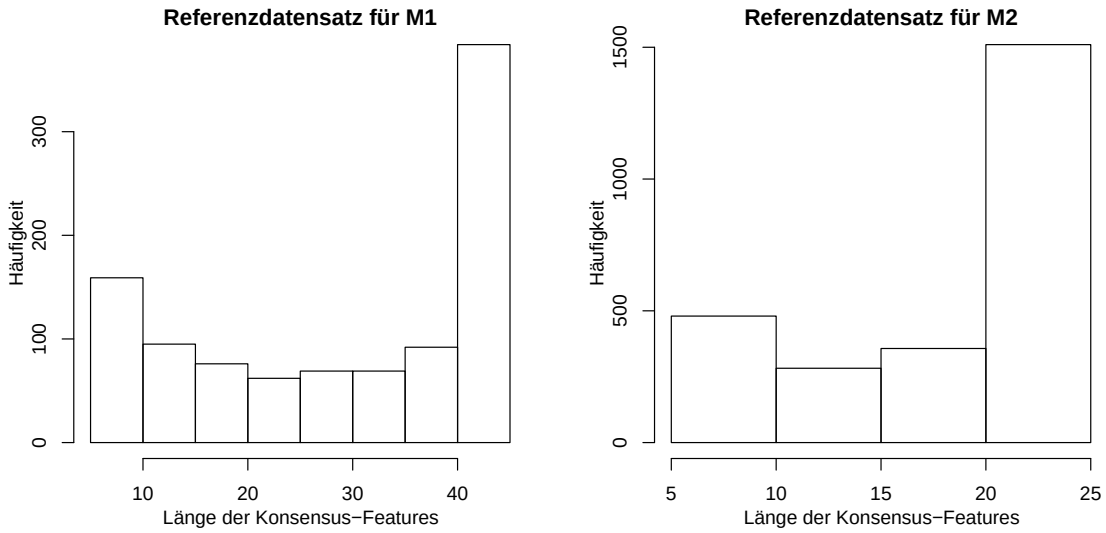

Abbildung 5.1: Verteilung der Länge der Konsensus-Features in den Referenzdatensätzen für M1 und M2. M1 enthält 44 und M2 24 LC/MS-Messungen.

Aufgrund der Auswahlkriterien erfasst der Referenzdatensatz nur einen kleinen Anteil der insgesamt vorhandenen Features. Im Datensatz M1 wurden für jede der 44 Messungen durchschnittlich ca. 6700 Features detektiert (Features insgesamt: 294791). Im

Referenzdatensatz für M1 befinden sich 2631 Konsensus-Elemente mit insgesamt 29728 Features. Der Datensatz M2 umfasst 24 Messungen, für die durchschnittlich 16122 Features detektiert wurden (Features insgesamt: 386920). Im erstellten Referenzdatensatz für M2 befinden sich 1008 Konsensus-Elemente mit insgesamt 49256 Features.

5.3.3 Charakteristik der Datensätze M1 & M2

Die Datensätze M1 und M2 stellen verschiedene Herausforderungen an die Alignment Algorithmen: M1 enthält relativ viele Messungen (44), aber dafür weniger Features pro Messung (durchschnittlich rund 6700) als M2. M2 umfasst „nur“ 24 Messungen, dafür sind die Feature-Listen mit durchschnittlich rund 16122 Features wesentlich dichter besetzt. Der Datensatz M1 zeigt, bedingt durch die ältere Technik, stärkere Abweichungen als M2, sowohl in der Retentionszeit als auch in m/z. Die durchschnittliche absolute Retentionszeitabweichung im Referenzdatensatz beträgt bei M1 5.4 s und bei M2 2.7 s. Die maximale Zeitdifferenz innerhalb eines Konsensus-Features beträgt bei M1 90 s und bei M2 20 s.
Generell lassen sich in beiden Datensätzen zwei verschiedene Arten von Abweichungen betrachten. Zum einen sind dies durch die Flußraten-Variabilität bei der Chromatographie verursachte systematische Retentionszeitabweichungen, die vor allem lokalen Charakter haben. So kann z.B. in einem Zeitbereich eine positive Verschiebung beobachtet werden und in einem anderen eine negative. Eine nichtlineare Retentionszeitkorrektur sollte diese Effekte besser korrigieren können als eine lineare Methode. Weiterhin treten auch zum Teil recht starke lokale Abweichungen auf, die keinem allgemeinen Trend zu unterliegen scheinen. Jeder Alignment-Algorithmus muss beide Arten von Variabilität berücksichtigen und gegebenenfalls kompensieren.
Abbildung 5.2 zeigt die zeitlichen Abweichungen im Referenzdatensatz für je drei zufällig gewählte Messungen aus M1 und M2. Deutlich zu erkennen sind die beschriebenen zwei Arten von Abweichungen. Die rot markierten Features einer Messung aus M1 im oberen Teil der Abbildung 5.2 weisen beispielsweise eine systematische Verschiebung von ca. 10 s gegen die anderen dargestellen Messungen auf. Die blau und schwarz markierten Features aus zwei anderen Messungen von M1 weisen dagegen in der gewählten Darstellung nur geringe systematische Verschiebungen auf, es treten vielmehr lokale und weniger systematische Abweichungen auf. Im Datensatz M2 wurde generell ein Überwiegen der lokalen Verschiebungen im Gegensatz zu globalen Trends beobachtet. Die im unteren Teil der Abbildung 5.2 dargestellen blau markierten Features aus einem Datensatz von M2 zeigen eine geringe nichtlineare systematische Verschiebung, bei den rot und schwarz markierten Features überwiegen jedoch starke lokale Abweichungen gegenüber einer geringen, im dargestellen Ausschnitt eher linearen, systematischen Abweichung. Abbildung 8.1 im Anhang (Seite 130) zeigt die zeitlichen Abweichungen in den Referenzdatensätzen von M1

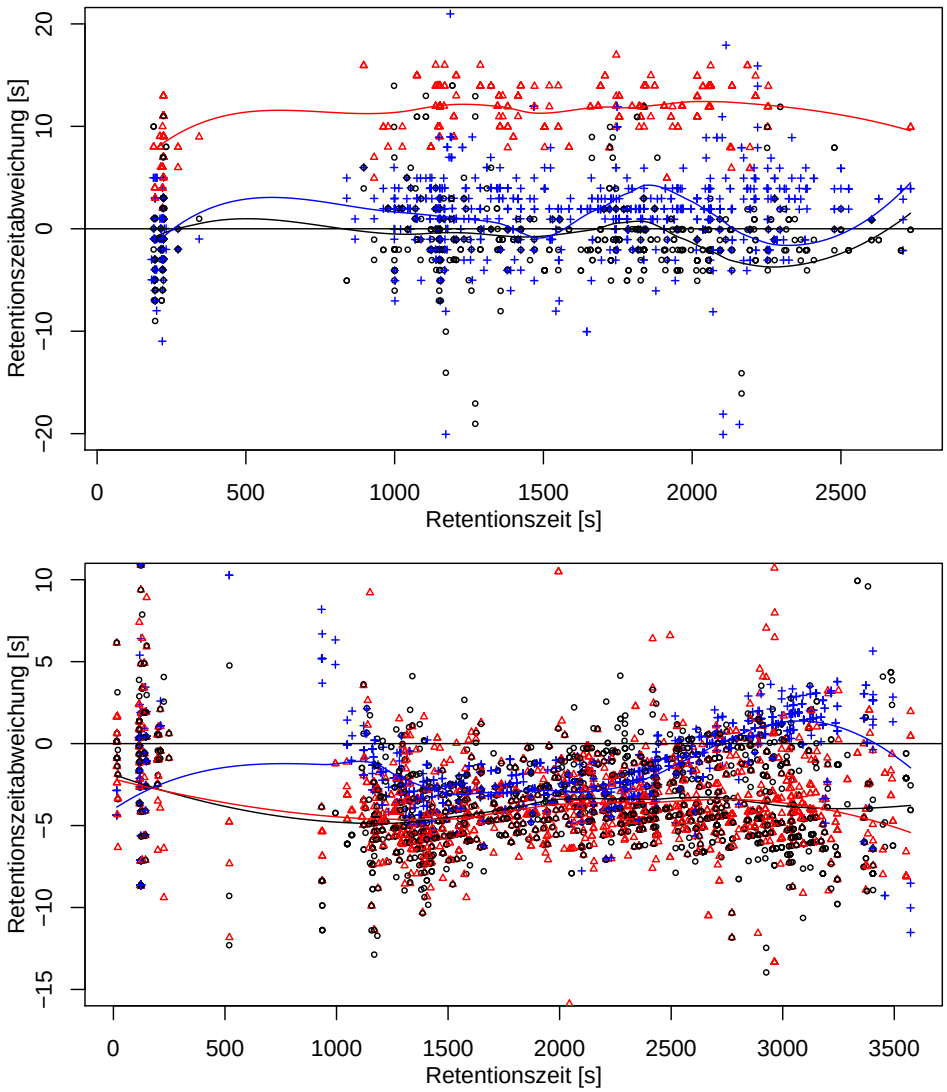

Abbildung 5.2: Retentionszeitabweichungen im Referenzdatensatz für je drei zufällig gewählte Messungen aus dem Datensatz M1 (oben) und dem Datensatz M2 (unten). Dargestellt sind jeweils die Differenzen der Retentionszeiten der Features aus Konsensus-Elementen zur Retentionszeit der entsprechenden Features aus der ersten Messung des Datensatzes. Loess-Regressionskurven wurden zur besseren Visualisierung hinzugefügt.

und M2 für die Konsensus-Features aus allen Messungen als Box-Whiskers-Darstellung.

5.4 Evaluierung

Betrachtet man ein bestimmtes Feature aus der Feature-Liste einer Messung als eine „Anfrage“ und die zurückgelieferten zugehörigen Features aus den Feature-Listen anderer Messungen als die „Antwort“ auf diese Frage, so lässt sich ein Alignment-Algorithmus bzw.

seine Ausgabe als ein Problem des Information Retrieval (IR) behandeln. Die Leistung eines IR Systems kann mittels der Maße *Precision* und *Recall* eingeschätzt werden (siehe auch Abschnitt 3.4.2). Die durchgeführte Evaluierung der Alignments orientiert sich an diesen Maßen und verwendet als Anpassung an das Alignment-Problem leicht modifizierte Definitionen von *Precision* und *Recall*.
Die von den einzelnen Programmen durchgeführte Retentionszeitkorrektur wird hier nur als Mittel zum Zweck der Bildung einer Konsensus-Liste gesehen und nicht getrennt bewertet. Die durchgeführte Bewertung bezieht sich ausschließlich auf die Konsensus-Liste bzw. die darin enthaltenen Konsensus-Features. Gefordert wird, für eine Anfrage genau die relevanten Features zu erhalten, ohne zusätzliche unerwünschte Antworten, was in diesem Fall falsch alignierten Features entspricht. Eine beim Alignment auftretende Besonderheit ist es, dass aus vielen Features bestehende Konsensus-Features unter Umständen vom Algorithmus nicht zusammenhängend gruppiert wurden und folglich in der vom Algorithmus erzeugten Konsensus-Liste als in kleinere Teilmengen „zerlegtes“ Konsensus-Features beobachtet werden. Da dieser Effekt natürlich nicht erwünscht ist, sollte ihm in der verwendeten Bewertung Rechnung getragen werden.

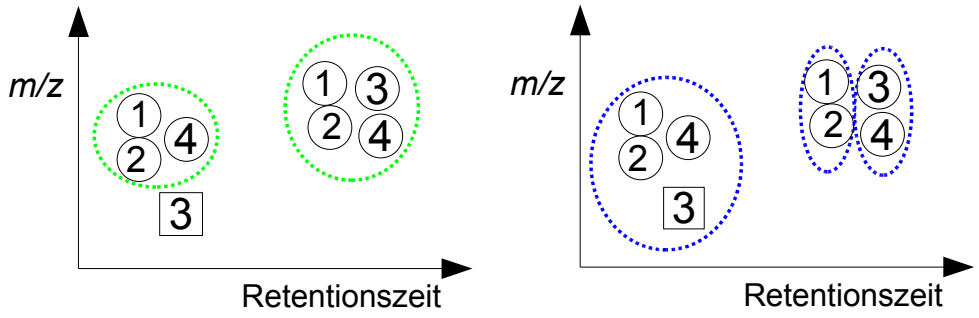

Abbildung 5.3: Beispiel für die Bewertung eines Alignments. In der linken Abbildung sind Features aus vier verschiedenen Feature-Listen (Markierung durch Zahlen) gezeigt. Die Zugehörigkeit zu zwei im Referenzdatensatz enthaltenen Konsensus-Features wurde durch farbige Ellipsen markiert. Die rechte Bildhälfte zeigt das durch einen Alignment-Algorithmus erhaltene Zuordnungs-Ergebnis. Das linke Konsensus-Feature enthält ein falsch zugeordnetes Feature, während das rechte vom Algorithmus in zwei Konsensus-Features zerlegt wurde. Der Alignment-Recall beträgt in diesem Beispiel 0.75, die Precision 0.875.

Es seien die Konsensus-Features im Referenzdatensatz mit ref_i bezeichnet, mit dem Index i, $i = 1, \ldots, N$. Die Konsensus-Features als Ergebnis eines Programms werden als tool_j bezeichnet, mit dem Index j, $j = 1, \ldots, M$. Betrachtet wird nun die Menge der vom Programm gelieferten Konsensus-Features und deren Schnittmenge mit einem gegebenen Konsensus-Feature aus dem Referenzdatensatz. Dann wird für jeden Index i mit M_i die Menge all der Indizes j bezeichnet, so dass $|\text{tool}_j| > 0$ und $|\text{ref}_i \cap \text{tool}_j| > 0$. Die

Kardinalität dieser Index-Menge, $|M_i|$, kann nun betrachtet werden als die Anzahl der Teile, in die ein Konsensus-Feature ref_i aus dem Referenzdatensatz im Ergebnis des Programms „zerlegt“ wurde. Die Vereinigungsmenge der Konsensus-Features tool_j bezüglich i sei $\widetilde{\text{tool}}_i := \bigcup_{j \in M_i} \text{tool}_j$. Dann ist $\widetilde{\text{tool}}_i$ die Menge aller Features, die erhalten wird, wenn alle Features aus ref_i in der vom Programm erhaltenen Konsensus-Liste „angefragt“ werden.
In Anlehnung an die klassische Definition von Precision und Recall, wird die *Alignment-Precision* definiert mit

$$\text{Precision}_{\text{Align}} = \frac{1}{N} \sum_{i=1}^{N} \frac{|\text{ref}_i \cap \widetilde{\text{tool}}_i|}{|\widetilde{\text{tool}}_i|} \tag{5.1}$$

und der *Alignment-Recall* als

$$\text{Recall}_{\text{Align}} = \frac{1}{N} \sum_{i=1}^{N} \frac{|\text{ref}_i \cap \widetilde{\text{tool}}_i|}{|M_i| \cdot |\text{ref}_i|} \,. \tag{5.2}$$

Der Faktor $|M_i|$ im Nenner von (5.2) dient als Strafterm für die Zerlegung eines Konsensus-Features aus dem Referenzdatensatz. Bei einem perfekten Alignment wären beide Maße gleich eins. Falsch positive (fehlerhaft alignierte) Features senken die Alignment-Precision, falsch negative (fehlerhaft nicht alignierte) Features senken den Alignment-Recall. Ein Beispiel zur Bewertung eines Alignments ist in Abbildung 5.3 gezeigt.
Zur automatischen Berechnung der Recall- und Precision-Werte wurde ein R-Skript benutzt. Die Laufzeiten der Programme wurden inklusive aller Dateioperationen gemessen. Durchgeführt wurden die Messungen auf einem AMD Athlon 64 X2 Dual Core Processor 4800+ mit 2GB RAM unter Linux (Ubuntu 6.06). Da XAlign nur unter Windows lauffähig ist, wurde es unter Windows XP evaluiert, welches in einer virtuellen Maschine unter VMWare Workstation 5.5.3 auf demselben Rechner lief (ein natives Windows XP sollte ca. 10-20% schneller sein).

5.4.1 Parameteroptimierung

Die Parameteroptimierung wurde unter besonderer Berücksichtigung des Recall-Wertes durchgeführt. Grund dafür ist zum einen der bei der Berechnung von $\text{Recall}_{\text{Align}}$ enthaltene Strafterm für die „Zerlegung“ von Konsensus-Features aus dem Referenzdatensatz. Eine solche Aufspaltung von eigentlich zusammengehörigen Features in mehrere Konsensus-Features ist in der Praxis besonders störend und erschwert die Auswertung. Zum anderen kann die Precision nicht ausreichend gut bewertet werden, da aufgrund der Auswahlkriterien die in einem Konsensus-Feature enthaltenen Features zwar mit hoher Sicherheit zusammengehören, jedoch nicht gänzlich ausgeschlossen werden kann, dass noch weitere Features dem Konsensus-Feature zuzuordnen sind. Ein Beispiel dafür können Features

sein, die in einigen Messungen eine etwas geringere Intensität aufwiesen und dadurch bei der Bildung der Feature-Gruppen durch die Korrelationsanalyse ausgeschlossen wurden, da der entsprechende Korrelationskoeffizient unter dem Schwellwert lag. In diesen Fällen wäre die Bewertung als falsch-positiv aligniertes Feature nicht zutreffend. Die durchgeführte Parameteroptimierung zeigte jedoch, dass in dem getesteten Parameterbereich der maximale erreichte Recall-Wert im Allgemeinen auch mit einem für das Programm maximalen Precision-Wert einhergeht.

Es wurden umfangreiche Testläufe durchgeführt, um die Parameter der einzelnen Programme zu optimieren, die vor allem die zulässige Toleranz in m/z und der Retentionszeit kontrollieren. Ausgehend von den bekannten maximalen Abweichungen der Datensätze wurden die Parameter der Programme innerhalb eines sinnvoll erscheinenden Bereichs für beide Datensätze variiert. Als Anhaltspunkt diente dafür ein Bereich von $0.01 \ldots 0.05$ m/z für den die m/z-Abweichung kontrollierenden Parameter (im Weiteren oft mit $\Delta m/z$ abgekürzt), sowie ein Bereich von $10 \ldots 100$ s für Parameter, welche die zulässige Retentionszeitabweichung (ΔRT) bestimmen. Eine zu erwartende Beobachtung war dabei, dass bei beiden Parametern zunächst ein bestimmter Mindestwert erreicht werden muss, um brauchbare Ergebnisse zu erhalten. Nicht unbedingt zu erwarten war jedoch, dass eine weitere Erhöhung dieser Werte (in einem sinnvollen Rahmen) in mehreren Fällen nahezu unveränderte Resultate lieferte.

MZmine

Bei MZmine trat dieser Effekt insbesondere beim Datensatz M2 auf. Ab einer zeitlichen Abweichung von 30 s und einer m/z-Abweichung von 0.02 wurden die für den Algorithmus maximalen Werte für Precision und Recall erreicht. Eine Erhöhung auf 50 s und mehr, sowie eine bis 0.05 erhöhte m/z-Abweichung änderte das Ergebnis nicht. Beim Datensatz M1, der höhere m/z- und Retentionszeitabweichungen aufweist, traten die höchsten beobachteten Recall-Werte erst ab $\Delta m/z$ =0.03 und ΔRT =50 s auf. Für größere $\Delta m/z$ und ΔRT sanken sowohl Recall als auch Precision. Für $\Delta m/z$ =0.02 und ΔRT =30 s wurde eine um ca. 1% höhere Precision beobachtet, wobei der Recall aber schon 5% niedriger als bei Maximum war.

MapAlignment (OpenMS)

Das Verhalten des MapAlignment-Algorithmus wird über eine Initialisierungs-Datei gesteuert, die neben den Bezeichnungen der Ein- und Ausgabedateien insgesamt 31 Parameter enthält. Die Einstellungen dieser Parameter wurden von der Autorin des Programms (E. Lange) vorgenommen, wobei sich die gewählten Parameterwerte für M1 und M2 nicht unterscheiden. Um die Auswirkungen von abweichenden Einstellungen zu testen, wurden die von der Autorin als einflussreich bezeichneten Parameter variiert. Dies betrifft die bei der Beschreibung des Algorithmus bereits erwähnten Parameter mz_bucket_size, $precision\_RT$ sowie $precision\_MZ$.

Eine Variation des Parameters *mz_bucket_size* zwischen *m/z* 0.01 und 0.1 ergab keine messbaren Unterschiede, ebenso eine probeweise Halbierung des Wertes von *precision_RT* von 100 s auf 50 s. Unterschiede ergaben dagegen veränderte Werte für *precision_MZ*. Für den Datensatz M1 wurden für Werte von *m/z* 0.1 bis 0.02 nur geringfügige Änderungen (<1%) in Precision und Recall festgestellt, für *m/z* 0.01 bei diesem Parameter sanken sowohl Recall (-5%) als auch Precision (-1%). Die Tendenzen für Werte < 0.1 *m/z* beim Datensatz M2 waren ähnlich. Probeweise wurden auch höhere Werte für diesen Parameter gewählt, was jedoch zu einer Verschlechterung der Ergebnisse führte (z.B. *m/z* 0.15: Recall -6%, Precision -2%). Die beobachteten Veränderungen sind jeweils als Differenz zu den als optimal ermittelten Recall- und Precision-Werten angegeben.

XAlign

Beim Programm XAlign waren die Beobachtungen bei der Optimierung ähnlich wie bei MZmine. Die höchsten Werte von Recall und Precision wurden für M1 bei $\Delta m/z$ =0.03 und ΔRT =30 s und für M2 bei $\Delta m/z$ =0.04 und ΔRT =30 s erreicht. Variierende Werte wie $\Delta m/z$ =0.02 oder 0.05 bzw ΔRT =20 oder 40 ergaben bei M2 lediglich Schwankungen der Precision im 1% Bereich. Für den Datensatz M1 lag der Recall-Wert für $\Delta m/z$ =0.02 dagegen schon um $\approx$ 4% niedriger, bei nahezu gleichbleibender Precision.

XCMS

Die Parameteroptimierung für XCMS wurde zunächst für die einfache Gruppierung ohne Anwendung der zusätzlichen Retentionszeitkorrektur durchgeführt. Die maximale Werte für Precision und Recall wurden für beide Datensätze mit den Parametern *bw*=30 sowie *mzwid*=0.05 beobachtet. Niedrigere Werte bei beiden Parametern ergaben sowohl schlechtere Recall- als auch Precision-Werte. Eine weitere Erhöhung des *bw*-Parameters, der die zeitliche Abweichung kontrolliert, ergab einen nahezu unveränderten Recall, jedoch eine geringe Precision, z.B. bei Datensatz M2 für *bw*=40 $\approx$4% niedriger. Der *minfrac*-Parameter wurde bei der Gruppierung auf 0.05 gesetzt, d.h. ein gültiges Konsensus-Features muss Features aus mindestens 5% der alignierten Messungen enthalten.

Zur Retentionszeitkorrektur wurde die nichtlineare Methode gewählt. Tests mit der linearen Korrekturmethode ergab in jedem Fall schlechtere Werte. Zur Anwendung der Retentionszeitkorrektur wurde zunächst die einfache Gruppierung mit den als optimal ermittelten Parameterwerten durchgeführt. Bei dieser ersten Gruppierung wurde der Parameter minfrac auf 1 gesetzt, d.h. es werden nur solche Konsensus-Features für die Retentionszeitkorrektur verwendet, die allen Feature-Listen vorkommen. Anschließend wurde die Retentionszeitkorrektur angewandt, wobei hierbei der höhere Glättungsparameter 0.75 statt der vorgegebenen 0.5 im Endergebnis bessere Werte lieferte. Nach diesem Schritt wurde erneut die Gruppierung durchgeführt, mit dem Unterschied, dass jetzt der bw-Parameter nur noch auf ein Drittel des Wertes bei der Erstgruppierung (d.h. bw=10) gesetzt wurde. Diese Vorgehensweise liefert erfahrungsgemäß die besten Ergebnisse und

wird auch vom Autor des Programms in ähnlicher Art empfohlen. Der minfrac-Parameter wurde bei der letzten Gruppierung auf 0.05 gesetzt.
Die Anwendung der Retentionszeitkorrektur ergab für den Datensatz M2 eine Verbesserung des Recalls im Vergleich mit den Werten bei der einfachen Gruppierung um 1%, der vorher verhältnismäßig schlechte Precision-Wert erhöhte sich dabei jedoch um ganze 20%. Anders beim Datensatz M1: Hier sank der Recall nach Anwendung der Korrektur um etwas mehr als 3%, die Precision erhöhte sich dabei um 7%.
Tabelle 5.1 zeigt die als Ergebnis der Parameteroptimierung gewählten Einstellungen.

Programm	Parameter	Datensatz M1	Datensatz M2
MZmine	m/z tolerance size	0.03	0.025
	RT tolerance size	50	30
OpenMS	m/z bucket	0.01	0.01
	precision m/z	0.1	0.1
	precision RT	100	100
XAlign	m/z variation	0.04	0.03
	RT variation	30	30
XCMS	mzwid	0.05	0.05
	bw	30	30
	span	0.75	0.75

Tabelle 5.1: Alignment-Parameter nach der Optimierung. Die Parameter für die Retentionszeitabweichungen sind in Sekunden angegeben.

5.4.2 Alignment Ergebnisse

Die Ergebnisse der Programme werden bezüglich der mit den optimierten Parametern erhaltenen $\text{Recall}_{\text{Align}}$ und $\text{Precision}_{\text{Align}}$ Werte diskutiert. Abbildung 5.4 zeigt alle Werte im Überblick. Diese sind nochmals in tabellarischer Form im Anhang (Tab.8.20, Seite 131) aufgeführt.

Datensatz M1

MZmine, OpenMS und XAlign erreichen auf dem Datensatz M1 vergleichbare Ergebnisse, mit durchschnittlichen Werten von 88% Recall und 71% Precision. Von diesen drei Programmen sind die Werte von MZmine mit 89% Recall und 74% Precision noch am höchsten, was insofern erstaunlich ist, da MZmine keine Retentionszeitkorrektur durchführt. Aufgrund der beobachteten, im Vergleich zu M2 deutlich höheren Retentionszeitabweichungen in diesem Datensatz wäre hier eher zu vermuten gewesen, dass die Ergebnisse

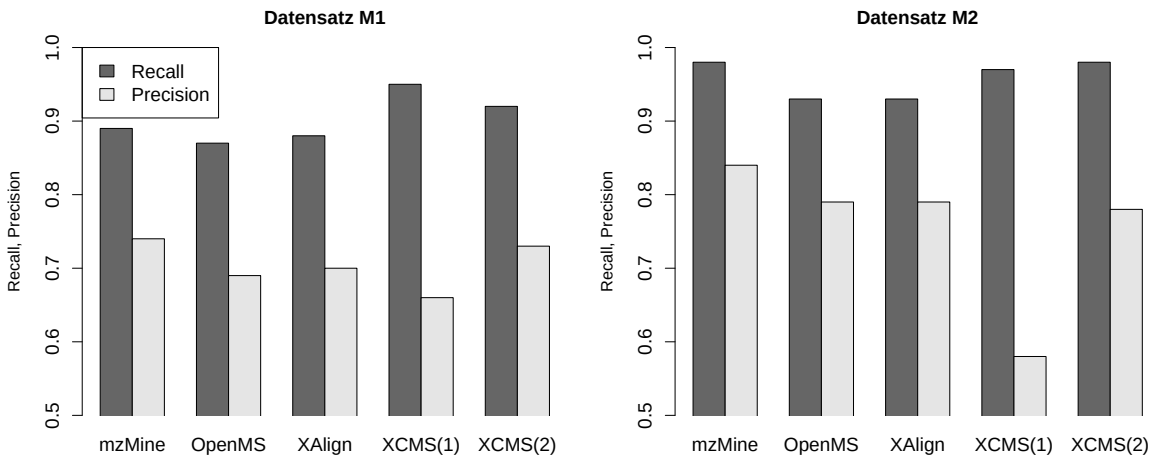

Abbildung 5.4: Alignment-Recall und -Precision Werte der getesteten Algorithmen auf den Datensätzen M1 und M2. XCMS wurde sowohl ohne (1) als auch mit (2) Retentionszeitkorrektur evaluiert.

von OpenMS und XAlign besser sind, da diese mehr Aufwand betreiben, um Retentionszeitabweichungen zu kompensieren.
XCMS erreicht auf diesem Datensatz mit 95% den besten Recall-Wert, erstaunlicherweise wird dieser jedoch *ohne* Retentionszeitkorrektur erreicht. Dafür ist jedoch die Precision mit 66% in dieser Einstellung der niedrigste beobachtete Wert. Wird die Retentionszeitkorrektur angewendet, so sinkt der Recall um 3%, die Precision steigt jedoch um 7%. Insofern ist nicht leicht zu entscheiden, ob die Retentionszeitkorrektur in diesem Fall als nutzbringend bewertet werden kann. Es scheint entweder die Korrektur ungenügend zu sein oder aber es sind noch viele lokale, nicht-systematische und daher nicht korrigierbare Abweichungen vorhanden, was den niedrigeren Recall-Wert nach der Retentionszeitkorrektur in Verbindung mit der erneuten Gruppierung bei geringeren Toleranzwerten erklären könnte.

Datensatz M2

Die erreichten Werte liegen auf dem Datensatz M2, der sowohl in der Retentionszeit als auch in m/z geringere Abweichungen aufweist, deutlich höher als bei M1. Der durchschnittliche Recall aller Programme beträgt hier 96%, die Precision 76%. OpenMS und XAlign zeigen hier die schlechtesten Recall-Werte mit jeweils 93%, bei einer Precision von ebenfalls jeweils 79%.
XCMS profitiert bei diesem Datensatz deutlich von der Retentionszeitkorrektur. So steigt nach der Anwendung der Recall zwar nur um 1% auf 98%, die Precision aber um 20% auf 78%. MZmine erzielt ebenfalls einen Recall von 98%, jedoch bei einer Precision von 84%, was damit das auf M2 insgesamt beste Ergebnis darstellt.

Laufzeit
XCMS benötigt für die komplette Berechnung beider Datensätze rund sieben Minuten, OpenMS 13 Minuten. MZmine und XAlign rechnen jedoch deutlich länger. So beträgt die gesamte Laufzeit für beide Datensätze bei MZmine 64 Minuten und bei XAlign 84 Minuten. Tabelle 8.21 im Anhang (Seite 131) zeigt die einzelnen Laufzeiten für jeden Datensatz.

5.5 Diskussion

XAlign und der MapAlignment Algorithmus von OpenMS erreichen im durchgeführten Vergleich insgesamt die schlechtesten Ergebnisse. Die Ursache für die unbefriedigenden Ergebnisse könnte bei beiden Algorithmen in der nur mittels einer linearen Regression durchgeführten Retentionszeitkorrektur liegen. Es ist denkbar, dass ein nichtlinearer Ansatz hier bessere Werte liefern würde. Ein Kritikpunkt an XAlign ist auch die im Vergleich zu den anderen Algorithmen besonders hohe Laufzeit. So dauerte das Alignment für den Datensatz M1 51 Minuten, was um einen Faktor von ≈50 höher ist als die Laufzeit beim schnellsten Programm (XCMS). Für den Datensatz M2 relativiert sich dieser Faktor auf ≈6. Eine relativ hohe Anzahl von zu alignierenden Messungen scheint demnach problematisch für die Laufzeit von XAlign zu sein.
MZmine benutzt von den verglichenen Programmen den einfachsten Algorithmus zum Alignment. Nichtsdestotrotz wurden damit auf dem Datensatz M2 die besten Ergebnisse erzielt. Beim Datensatz M2 lag das Ergebnis nur im Mittelfeld, was auf die fehlende Retentionszeitkorrektur zurückgeführt werden könnte. MZmine hat die nach XAlign zweitlängste Laufzeit, 20 Minuten auf M1 und 44 Minuten auf M2. Für die Laufzeit von MZmine scheint daher eher die Größe der Feature-Listen entscheidend zu sein. Die Ursache hierfür dürften die wiederholte Iterationen über die einzelnen Feature-Listen bei der Erstellung der *master feature list* sein, was mit einem optimierten Algorithmus zum Auffinden der jeweils benötigten benachbarten Features verbessert werden könnte.
XCMS erreichte auf dem Datensatz M1 den insgesamt höchsten und auf M2 gemeinsam mit MZmine den höchsten Recall-Wert. Auf M1 wurde dieser jedoch ohne Retentionszeitkorrektur erreicht und auf M2 mit durchgeführter Korrektur. Da sich der Recall wie hier auf M1 gezeigt, nach Anwendung der Retentionszeitkorrektur auch verschlechtern kann, ist es für den Anwender somit schwer einzuschätzen, welche Methode angewandt werden sollte, wenn ein hoher Recall-Wert von Bedeutung ist. In Hinsicht auf die nach Anwendung der Retentionszeitkorrektur jeweils deutlich verbesserten Precision-Werte scheint diese jedoch im Allgemeinen vorteilhaft. XCMS zeigte die im Vergleich kürzesten Laufzeiten für die Alignments.
Generell scheinen Reserven in der Leistungsfähigkeit der Alignment-Algorithmen zu bestehen. Denkbar wäre beispielsweise, die Ideen aus gut funktionierenden Ansätzen zu

kombinieren, z.B. den Gruppierungsalgorithmus aus MZmine mit einer nichtlinearen Retentionszeitkorrektur wie der von XCMS zu verbinden. Weitere Möglichkeiten bestehen in der Kombination mit rohdatenbasierten Ansätzen zur Retentionszeitkorrektur wie beispielsweise OBI-Warp [Prince und Marcotte 2006]. Der Vorteil bei solchen Methoden besteht darin, dass keine vorherige Gruppierung zur Findung von „landmark“-Features für die Retentionszeitkorrektur nötig ist. Die Retentionszeitabweichungen werden direkt auf den Rohdaten geschätzt, z.B. bei OBI-Warp über einen Dynamic Time Warping Ansatz auf einer Matrix, welche die Korrelationswerte der einzelnen Massenspektren von zwei LC/MS Messungen enthält. Ein derartiger Ansatz könnte dann erfolgversprechend sein, wenn sich die Suche nach „landmark“-Features sehr schwierig gestaltet. Dies wird vor allem bei LC/MS Datensätzen beobachtet, die in größerem zeitlichen Abstand gemessenen wurden und dadurch erhebliche Retentionszeitverschiebungen aufweisen. Eine so erhaltene Retentionszeitkorrektur sollte die dann durchgeführte Gruppierung der Feature-Listen erheblich vereinfachen.

5.6 Zusammenfassung

Mithilfe der in Abschnitt 4.3.2 beschriebenen korrelationsbasierten Methode zur Gruppierung der von einer Substanz hervorgerufenen Features wurde ein Referenzdatensatz erstellt, um damit erstmalig eine Evaluierung von Alignment-Algorithmen auf Metabolomik LC/MS Daten durchzuführen.
Keines der verglichenen Programme lieferte Ergebnisse, die als unzureichend bezeichnet werden müssten, jedoch brachte der Vergleich auch keinen Algorithmus als klaren Favoriten hervor. Ähnliches gilt im Übrigen auch für den Proteomik-Bereich, siehe [Lange et al. 2008].
Die beiden für Metabolomik konzipierten Programme, MZmine und XCMS, zeigten im Vergleich mit den anderen Programmen die besten $\text{Recall}_{\text{Align}}$-Werte. XCMS zeichnete sich zwar durch die kürzeste Laufzeit aus, lieferte aber niedrigere $\text{Precision}_{\text{Align}}$-Werte als MZmine. MZmine führt dagegen keine Retentionszeitkorrektur durch und lieferte auf dem Datensatz mit stärkeren Zeitabweichungen schlechtere Werte als XCMS.
Die für Proteomik Daten entwickelten Programme OpenMS und XAlign erscheinen in der vorliegenden Version für das Alignment von Metabolomik Daten nur bedingt geeignet.

6 Zusammenfassung und Ausblick

Die Metabolomik hat sich innerhalb des letzten Jahrzehnts zu einer wichtigen Methode für die Erforschung und den Vergleich von Metaboliten in biologischen Systemen entwickelt. Als wesentliche Techniken werden dabei GC/MS, CE/MS, LC/MS und NMR genutzt. Die hier betrachtete Methode der HPLC/ESI-QTOF-MS zur Analyse von pflanzlichen Extrakten ist in der Lage, einen breiten Bereich (hinsichtlich der Molekülmasse und Polarität) von Sekundärmetaboliten zu messen.
In dieser Arbeit wurden drei wichtige Schritte bei der Verarbeitung von Metabolomik LC/MS Daten betrachtet: die Feature-Detektion, die Annotation zusammengehöriger Features sowie das Alignment von Feature-Listen aus mehreren Messungen.
Die Feature-Detektion, deren Aufgabe die Erkennung und Quantifizierung aller in einer LC/MS Messung enthaltenen Features ist, stellt dabei den entscheidenden Ausgangspunkt dar. Da die zu diesem Zeitpunkt verfügbaren Algorithmen zur Feature-Detektion von Metabolomik LC/MS Daten –MZmine und der matchedFilter-Algorithmus von XCMS– einige Defizite aufweisen, wurde ein neues Verfahren „centWave“ entwickelt.
Der centWave-Algorithmus verwendet dabei zunächst einen dichteorientierten Algorithmus zur Erkennung von potentiell interessanten Massensignalen, die als „regions of interest“ (ROI) bezeichnet werden. Der Vorteil hierbei ist, dass die ansonsten häufig verwendete Methode des „Binning“ zur Diskretisierung der Daten entfällt, womit auch einige der damit verbundenen Probleme umgangen werden.
Zur genaueren Untersuchung der Intensitätswerte in den detektierten ROI kommt in der zweiten Phase des centWave-Algorithmus eine auf Wavelets basierende Technik zum Einsatz. Die hierfür benutzte kontinuierliche Wavelet Transformation auf mehreren Skalen mit dem „Mexican Hat“-Wavelet ist in der Lage, chromatographische Peaks verschiedener Breite zuverlässig zu detektieren. Durch eine lokale Bestimmung von chromatographischer Basislinie und Rauschpegel können auch Features mit geringer Intensität –hervorgerufen durch Substanzen sehr niedriger Abundanz oder geringer Ionisierbarkeit– erkannt werden.
Das entwickelte Verfahren ist weitestgehend universell und wurde erfolgreich auf den Messungen von verschieden LC/MS Gerätekombinationen, wie HPLC/QTOF, UPLC/QTOF und HPLC/Orbitrap getestet.
Zum Vergleich mit den zwei anderen Feature-Detektions Algorithmen (matchedFilter und MZmine) wurde eine Evaluierung auf Verdünnungsreihen und Mischungen von *A.thaliana* Samen- und Blattextrakten durchgeführt. Gemessen wurden die Precision-, Recall- und F-score-Werte in Bezug auf einen vorher erstellten Referenzdatensatz. Die vom centWave-Algorithmus erzielten Recall-Werte lagen dabei im Mittel 14% und 4% höher als die von matchedFilter bzw. MZmine, der über Precision und Recall ermittelte F-score war durchschnittlich 9% bzw. 4% höher. Der centWave-Algorithmus liefert damit eine im Vergleich zu den anderen beiden Algorithmen deutlich verbesserte Feature-Detektion.

Die manuelle Interpretation der nach der Feature-Detektion erhaltenen, teilweise mehrere tausend Features umfassenden Listen, ist sehr zeitaufwendig. Aufgrund des Ionisierungsprozesses wird jede Verbindung durch eine Vielzahl von Features repräsentiert, aus denen unter Kenntnis der auftretenden Ionenbildung die Molekülmasse der Substanz abgeleitet werden muss. Um diesen Prozess zumindest teilweise zu automatisieren, wurden Methoden zur Gruppierung von zusammengehörigen Features und der darauf folgenden Erkennung der Molekülmasse der entsprechenden Verbindung erstellt. Diese Informationen können für eine kompakte Repräsentation des Datensatzes genutzt werden und erleichtern weitere Schritte der Interpretation, wie die Massendekomposition oder die Molekülsuche in Datenbanken.

Die Methode zur Feature-Gruppierung nutzt den Umstand, dass die chromatographischen Intensitäten der Features, die zu einer Verbindung gehören, in einem linearen Zusammenhang stehen. Verwendet wird daher eine korrelationsbasierte Technik zur Gruppierung dieser Features. Um eine Fehlzuordung der Features bei koeluierender Substanzen zu verhindern, wird auf den nach der Korrelationsanalyse erstellten Korrelationsgraphen eine Zerlegung in stark verbundene Subgraphen durchgeführt.

Zur Erkennung der Molekülmasse einer Verbindung wird ein regelbasiertes Verfahren genutzt, wobei die dafür verwendeten Regeln der Ionenbildung vom Benutzer angepasst werden können. Das Verfahren ist mit diesen Regeln in der Lage, die bei der Ionisierung gebildeten Addukte und Fragmente zu erkennen und aus diesen die Molekülmasse der Substanz abzuleiten. Weiterhin werden auch die in regelmäßigen Abständen auftretenden Isotopomere erkannt und zur Feststellung des Ladungszustandes der entsprechenden Ionen genutzt.

Um die Leistungsfähigkeit dieser Methode zu testen, wurde eine LC/MS Messung von 14 bekannten Verbindungen verwendet, für die eine manuelle Annotation vieler Features gegeben war. Die korrelationsbasierte Technik zur Gruppierung ergab für 84 von 91 Features eine richtige Gruppierung der zu einer Verbindung gehörenden Features, die Ausnahmen bildeten Features mit besonders niedriger Intensität. Die mithilfe der vorgegebenen Ionisierungsregeln durchgeführte Erkennung der Molekülmasse ergab in 11 von 14 Fällen ein richtiges Ergebnis.

Zur weiteren Verbesserung des Verfahrens sind umfangreiche Messungen von bekannten Verbindungen nötig, mit deren Hilfe umfassendere Regelmengen erstellt werden können, insbesondere für die momentan kaum erfassten Fragmentierungen von Molekülen. Eine weitere mögliche Entwicklungsrichtung ist die kombinierte Auswertung von Messungen derselben Probe im positiven und negativen Ionisierungsmodus. Da etliche Verbindungen bevorzugt in einem Modus ionisiert werden und bei den Ionisierungsmodi völlig unterschiedliche Ionen erzeugt werden, können die beiden resultierenden Feature-Listen momentan nur aufwendig manuell verglichen werden. Über eine zeitlich basierende Zuord-

nung der Feature-Gruppen aus beiden Ionisierungsmodi, in Verbindung mit einer entsprechend erweiterten Regelmenge unter Beachtung der Polarität, könnte damit eine Erkennung der Molekülmasse durchgeführt werden, die die Informationen aus beiden Ionisierungsmodi nutzt. Im Idealfall würde damit auch eine einzige –um die durch die Ionisierungsmodi auftretenden Redundanzen bereinigte– Featureliste als Ergebnis erzeugt.

Für viele Anwendungen im Metabolomik-Bereich ist es erforderlich, die Intensitäten von gemessenen Substanzen über mehrere, manchmal hunderte, LC/MS-Messungen hinweg zu vergleichen. Um die Vergleichbarkeit der Messungen zu ermöglichen ist es nötig, die auftretenden zeitlichen Abweichungen bei den chromatographischen Trennungen und auch der m/z-Werte zu berücksichtigen, bzw. gegebenenfalls zu kompensieren. Für dieses Alignment-Problem sind bereits eine Vielzahl von Programmen verfügbar, von denen jedoch ein großer Teil speziell für Proteomik-Daten entwickelt wurde. Um die Anwendbarkeit dieser Programme auf Metabolomik-Daten zu prüfen und die Ergebnisse der einzelne Programme zu vergleichen, wurde eine entsprechende Evaluierung durchgeführt. Für Metabolomik LC/MS-Daten existieren im Gegensatz zu Proteomik-Daten momentan noch keine Datenbanken zur umfangreichen Identifizierung von Features, welche zur Erstellung von Referenzdatensätzen benutzt werden könnten. Um in Ermangelung einer vollständigen chemischen Identifizierung trotzdem eine Evaluierung zu ermöglichen, wurde mithilfe der korrelationsbasierten Feature-Gruppierung eine Zuordnung von Feature-Gruppen über viele Messungen hinweg durchgeführt, womit schließlich ein Referenzdatensatz von sicher zugeordneten Features erstellt wurde.
Zur Auswertung wurden auf das Alignment-Problem angepasste Precision- und Recall-Maße genutzt. Betrachtet wurden die vier Programme MZmine, XCMS, OpenMS und XAlign auf zwei Datensätzen mit verschiedenen Abweichungen in der Retentionszeit und m/z. Die beiden für Metabolomik Daten entwickelten Programme, MZmine und XCMS, zeigten im Vergleich mit den anderen Programmen insgesamt die besten Recall-Werte, wobei bei XCMS etwas niedrigere Precision-Werte als bei MZmine gemessen wurden. MZmine führt im Gegensatz zu XCMS keine Retentionszeitkorrektur durch und lieferte daher auf dem Datensatz mit stärkeren Zeitabweichungen deutlich niedrigere Werte als XCMS. Die für Proteomik Daten entwickelten Programme OpenMS und XAlign zeigten auf beiden Datensätzen niedrigere Recall-Werte als XCMS und MZmine und erscheinen daher in der vorliegenden Version für das Alignment von Metabolomik Daten nur bedingt geeignet. Ursache dürfte bei beiden Programmen die nur mittels einer linearen Regression durchgeführte Retentionszeitkorrektur sein, die für die Korrektur der bei den Metabolomik-Daten auftretenden nichtlinearen Retentionszeitsverschiebungen nicht ausreicht.
Obwohl XCMS und MZmine bei der Evaluierung verhältnismäßig gute Werte lieferten, wurde jedoch kein vollständig zufriedenstellendes Ergebnis erreicht. Eine weitere Verbes-

serung der Algorithmen scheint möglich. Vorstellbare Entwicklungen sind hier beispielsweise die Kombination von rohdatenbasierten Ansätzen zur Retentionszeitkorrektur wie OBI-Warp [Prince und Marcotte 2006] mit dem Alignment auf Feature-Ebene.

Zusammenfassend lässt sich sagen, dass diese Arbeit einen wichtigen Beitrag zu einer verlässlichen, automatisierten Verarbeitung von Metabolomik LC/MS-Daten darstellt.
Das noch junge Forschungsgebiet der Metabolomik ermöglicht über die Messung einer großen Zahl von Metaboliten Aussagen über den funktionalen Status eines biologischen Systems. Aktuelle Anwendungen liegen z.B. in der Systembiologie, zusammen mit der funktionalen Genomik zur Annotation von Genen und Aufdeckung von Stoffwechselwegen. Die Systembiologie nutzt die kombinierte Messung und Interpretation von Metaboliten, Proteinen und/oder der Genexpression, um zu einem umfassenden Verständnis der dynamischen Interaktion der Komponenten in einem biologischen System zu gelangen.

7 Glossar

CE/MS
Kapillarelektrophorese (Capillary Electrophoresis) gekoppelt mit Massenspektrometrie

EIC
Extrahiertes Ionen Chromatogramm, Summierung der in einem definierten, schmalen Fenster in m/z auftretenden Massensignale in einem LC/MS Datensatz auf die Zeitachse

ESI
Elektrospray Ionisierung

FT-ICR
Fouriertransformations-Ionenzyklotronresonanz-Massenspektrometrie (Fourier transform ion cyclotron resonance mass spectrometry)

GC/MS
Gaschromatographie gekoppelt mit Massenspektrometrie

HPLC
Hochleistungsflüssigchromatographie

LC/MS
Flüssigchromatographie (i.A. HPLC) gekoppelt mit Massenspektrometrie

UPLC
„Ultra Performance Liquid Chromatography“, Variante der HPLC von der Firma Waters. Verwendet kleinere Partikelgrößen ($\approx 1.7\ \mu m$) und höheren Druck (bis 1000 bar) als die herkömmliche HPLC ($\approx 3 - 5\ \mu m$, 300 bar).

TIC
Total Ionen Chromatogramm, Summierung aller auftretenden Massensignale in einem LC/MS Datensatz auf die Zeitachse

TOF
Time-of-Flight Massenspektrometer

8 Anhang

Beschreibungen der Datensätze

MM14

Mischung der Substanzen o-Anissäure, Biochanin A, Ferulsäure, p-Coumarsäure, *N*-(3-Indolylacetyl)-L-Valine, Kinetin, Indole-3-Acetonitrile, Indole-3-Carbaldehyd, Kaempferol, Phloretin, Phlorizin, Phenylglycin, Rutin und Phenylalanin-d5 bei einer Konzentration von 20 μM. Die chromatographische Trennung wurde auf einem Acquity UPLC System (Waters) mit einer modifizierten C_{18} Säule und einem 20 min Wasser/Acetonitril Gradienten durchgeführt. Die eluierenden Substanzen wurden mittels eines Bruker MicrOTOF-Q Massenspektrometers im positiven Ionisierungsmodus bei einer Scan-Rate von 3 Hz detektiert. Die Massenkalibration wurde gegen Lithum-Formiat durchgeführt.

LSMIX

Mischungen bestehend aus Lösungsmittel, Samen- und Blattextrakten von *Arabidopsis thaliana* wurden in den folgenden Volumenverhältnissen angefertigt: 0/100/0, 25/75/0, 50/50/0, 75/25/0, 0/0/100, 25/0/75, 50/0/50, 75/0/25, 0/75/25, 0/50/50, 0/25/75, 100/0/0 (Lösungsmittel/Samen/Blatt). Die 12 Proben wurden in jeweils zehn technischen Replikaten gemessen. Chromatographische Trennung und Massenspektrometrie wie bei MM14 beschrieben.

E105

Messungen von *Arabidopsis thaliana* Blattextrakten. Die chromatographische Separation wurde auf einem HPLC-System von Agilent durchgeführt. Die eluierenden Substanzen wurden von 100 bis 1000 *m/z* mittels eines MicrOTOF-Q (Bruker Daltonics) Massenspektrometers detektiert, betrieben mit einer „Apollo II“ Elektrospray-Ionenquelle im positiven Ionisierungsmodus. Die Akkumulationszeit eines Scans betrug 1,5 Sekunden. Die Massenauflösung für ein Kalibrierungs-Peptid (bei *m/z* 829) betrug 14000.

E170

Messungen von Blattextrakten einer für die Infektion mit *Phytophthora infestans* empfindlichen *Arabidopsis thaliana* Mutante. Chromatographische Trennung und Massenspektrometrie wie bei MM14 beschrieben.

Tabelle zur Parameteroptimierung der Feature-Detektions Algorithmen

Algorithmus	Parameter	Start	Ende	Schrittweite	Ergebnis A	Ergebnis B
centWave	prefilter	(2,100)	(2,500)	(0,100)	(2,400)	(2,200)
	peakwidth	(5,10)	-	-	(5,10)	(5,10)
	ppm	20	30	2	30	30
	snthresh	2	10	1	5	4
matchedFilter	fwhm	2	6	1	4	4
	snthresh	2	15	1	12	7
	step	0.01	0.05	0.005	0.02	0.025
	max	10	100	10	50	50
MZmine	bin size	0.01	0.05	0.005	0.05	0.05
	chromatographic threshold level	0.50	0.95	0.05	0.7	0.85
	intensity tolerance	0.1	0.9	0.1	0.7	0.7
	minimum peak duration	3	6	1	3	3
	minimum peak height	0	1000	50	500	300
	m/z tolerance	0.01	0.05	0.01	0.03	0.03
	noise level	0	100	10	20	20

Tabelle 8.1: Parameter, Bereiche und Schrittweiten für die der Parameterdurchlauf zur Einstellung der Feature-Detektions-Algorithmen durchgeführt wurde, sowie die als Ergebnis der Optimierung gewählten Parameter.

Precision, Recall und F-score Werte zur Evaluierung der Feature-Detektions-Algorithmen

Experiment 1 : Evaluierung auf Verdünnungsreihen

Parametersatz A

Angegeben sind jeweils durchschnittliche Werte (in Prozent) für die Messungen auf zehn technischen Replikaten, sowie die Standardabweichung.

Konzentrationsverhältnis (%)			Precision (%)		
Lösungsmittel	Samen	Blatt	centWave	matchedFilter	MZmine
75	25	0	75.9 ± 1.4	84.1 ± 1	52.4 ± 2.6
50	50	0	79.8 ± 1	88.5 ± 0.6	79.9 ± 0.7
25	75	0	81.5 ± 0.8	90.4 ± 0.7	81.5 ± 0.6
75	0	25	79 ± 1.5	70.7 ± 1	77.9 ± 1.5
50	0	50	80.6 ± 1.1	73.2 ± 0.7	74.5 ± 1.1
25	0	75	78.7 ± 1	73 ± 0.8	70.6 ± 1

Tabelle 8.2: Precision-Werte für samen- (oberer Teil) und blattspezifische (unterer Teil) Features aus dem Referenzdatensatz A.

Konzentrationsverhältnis (%)			Recall (%)		
Lösungsmittel	Samen	Blatt	centWave	matchedFilter	MZmine
75	25	0	49.9 ± 0.4	40.6 ± 0.6	56.3 ± 1.1
50	50	0	71.8 ± 0.6	52.8 ± 0.4	67.2 ± 0.6
25	75	0	90 ± 1.2	62 ± 0.6	84.8 ± 1.4
75	0	25	52 ± 0.7	43.9 ± 0.7	50.1 ± 0.7
50	0	50	73.2 ± 0.9	56.3 ± 0.6	69.8 ± 0.5
25	0	75	88.2 ± 0.9	65.1 ± 0.8	84.7 ± 0.8

Tabelle 8.3: Recall-Werte für samen- (oberer Teil) und blattspezifische (unterer Teil) Features aus dem Referenzdatensatz A.

Konzentrationsverhältnis (%)			F-score (%)		
Lösungsmittel	Samen	Blatt	centWave	matchedFilter	MZmine
75	25	0	60.2 ± 0.5	54.8 ± 0.7	54.3 ± 1.9
50	50	0	75.6 ± 0.7	66.2 ± 0.4	73 ± 0.5
25	75	0	85.5 ± 0.7	73.6 ± 0.6	83.1 ± 0.7
75	0	25	62.7 ± 0.7	54.2 ± 0.8	61 ± 0.7
50	0	50	76.7 ± 0.5	63.6 ± 0.6	72.1 ± 0.6
25	0	75	83.1 ± 0.6	68.8 ± 0.7	77.1 ± 0.8

Tabelle 8.4: F-score für samen- (oberer Teil) und blattspezifische (unterer Teil) Features aus dem Referenzdatensatz A. Der F-score wurde berechnet mittels der Recall-Werte in Tabelle 8.3 und der Precision-Werte in Tabelle 8.2.

Parametersatz B

Konzentrationsverhältnis (%)			Precision-Werte (%)		
Lösungsmittel	Samen	Blatt	centWave	matchedFilter	MZmine
75	25	0	73.2 ± 1.2	78.7 ± 1	71.1 ± 2.8
50	50	0	75.5 ± 0.8	83.1 ± 1	74.6 ± 0.8
25	75	0	73.7 ± 1	84.9 ± 1.1	75.2 ± 0.9
75	0	25	71 ± 0.9	56.3 ± 1	69.4 ± 2.1
50	0	50	72.2 ± 0.8	57.7 ± 0.7	69.2 ± 1.4
25	0	75	67.4 ± 0.7	56.9 ± 0.9	65.7 ± 1.5

Tabelle 8.5: Precision-Werte für samen- (oberer Teil) und blattspezifische (unterer Teil) Features aus dem Referenzdatensatz B.

Konzentrationsverhältnis (%)			Recall (%)		
Lösungsmittel	Samen	Blatt	centWave	matchedFilter	MZmine
75	25	0	52.6 ± 0.6	48.7 ± 0.6	45.5 ± 0.8
50	50	0	73.7 ± 0.7	64.4 ± 0.5	65.1 ± 0.6
25	75	0	87.7 ± 1.1	75.5 ± 0.6	80.2 ± 1.1
75	0	25	53.5 ± 0.8	49.4 ± 0.9	49.7 ± 0.7
50	0	50	73.4 ± 0.9	65.1 ± 0.5	69.1 ± 0.6
25	0	75	84.1 ± 1	75.8 ± 0.9	80.9 ± 0.8

Tabelle 8.6: Recall-Werte für samen- (oberer Teil) und blattspezifische (unterer Teil) Features aus dem Referenzdatensatz B.

Konzentrationsverhältnis (%)			F-score (%)		
Lösungsmittel	Samen	Blatt	centWave	matchedFilter	MZmine
75	25	0	61.2 ± 0.7	60.2 ± 0.7	55.5 ± 1.3
50	50	0	74.6 ± 0.7	72.6 ± 0.6	69.5 ± 0.5
25	75	0	80.1 ± 0.8	79.9 ± 0.7	77.6 ± 0.8
75	0	25	61 ± 0.7	52.6 ± 1	57.9 ± 1
50	0	50	72.8 ± 0.7	61.2 ± 0.6	69.1 ± 0.9
25	0	75	74.8 ± 0.5	65 ± 0.9	72.5 ± 1.2

Tabelle 8.7: F-score für samen- (oberer Teil) und blattspezifische (unterer Teil) Features aus dem Referenzdatensatz B. Der F-score wurde berechnet mittels der Recall-Werte in Tabelle 8.6 und der Precision-Werte in Tabelle 8.5.

Experiment 2 : Evaluierung auf komplexen Mischungen

Parametersatz A

Konzentrationsverhältnis (%)			Precision (%)		
Lösungsmittel	Samen	Blatt	centWave	matchedFilter	MZmine
0	25	75	79.9 ± 0.9	77.5 ± 1.1	73.5 ± 1.1
0	50	50	80.9 ± 1.3	81.3 ± 0.9	76.8 ± 1
0	75	25	81.7 ± 0.8	85.1 ± 1	80.7 ± 0.9

Tabelle 8.8: Precision-Werte für die Vereinigungsmenge von samen- und blattspezifischen Features aus dem Referenzdatensatz A.

Konzentrationsverhältnis (%)			Recall (%)		
Lösungsmittel	Samen	Blatt	centWave	matchedFilter	MZmine
0	25	75	61.5 ± 1	42.6 ± 0.5	58.7 ± 0.7
0	50	50	62.5 ± 0.7	42.7 ± 0.7	60 ± 0.8
0	75	25	62.1 ± 1.1	41.7 ± 0.7	59.5 ± 1.1

Tabelle 8.9: Recall-Werte für die Vereinigungsmenge von samen- und blattspezifischen Features aus dem Referenzdatensatz A.

Konzentrationsverhältnis (%)			F-score (%)		
Lösungsmittel	Samen	Blatt	centWave	matchedFilter	MZmine
0	25	75	69.5 ± 1	55 ± 0.6	65.3 ± 0.7
0	50	50	70.5 ± 0.9	56 ± 0.7	67.4 ± 0.8
0	75	25	70.6 ± 0.8	56 ± 0.5	68.5 ± 0.8

Tabelle 8.10: F-score für die Vereinigungsmenge von samen- und blattspezifischen Features aus dem Referenzdatensatz A. Der F-score wurde berechnet mittels der Recall-Werte in Tabelle 8.9 und der Precision-Werte in Tabelle 8.8.

Konzentrationsverhältnis (%)			Recall (%)		
Lösungsmittel	Samen	Blatt	centWave	matchedFilter	MZmine
0	25	75	47.6 ± 1.1	32.8 ± 0.7	45.1 ± 0.9
0	50	50	63.6 ± 1.5	41.9 ± 1.3	60.7 ± 1.4
0	75	25	79 ± 2.4	50.4 ± 1.8	75.7 ± 2.3
0	75	25	53.4 ± 0.9	40.3 ± 1	51.6 ± 0.7
0	50	50	69.9 ± 0.6	51.1 ± 0.8	68 ± 0.8
0	25	75	84.6 ± 1.1	60.7 ± 0.4	81.7 ± 0.5

Tabelle 8.11: Recall-Werte für samen- (oberer Teil) und blattspezifische (unterer Teil) Features aus dem Referenzdatensatz A.

Parametersatz B

Konzentrationsverhältnis (%)			Precision (%)		
Lösungsmittel	Samen	Blatt	centWave	matchedFilter	MZmine
0	25	75	72.3 ± 1.3	63.6 ± 1.5	69.7 ± 1.7
0	50	50	74.8 ± 0.8	69.3 ± 0.8	72.2 ± 1.3
0	75	25	74.6 ± 0.7	74.2 ± 0.6	75.1 ± 1.1

Tabelle 8.12: Precision-Werte für die Vereinigungsmenge von samen- und blattspezifischen Features aus dem Referenzdatensatz B.

Konzentrationsverhältnis (%)			Recall (%)		
Lösungsmittel	Samen	Blatt	centWave	matchedFilter	MZmine
0	25	75	59.8 ± 0.9	49.7 ± 1	55 ± 0.7
0	50	50	62 ± 0.8	50.3 ± 0.7	56.6 ± 0.7
0	75	25	61 ± 1.1	49.1 ± 1	56.2 ± 0.9

Tabelle 8.13: Recall-Werte für die Vereinigungsmenge von samen- und blattspezifischen Features aus dem Referenzdatensatz B.

Konzentrationsverhältnis (%)			F-score (%)		
Lösungsmittel	Samen	Blatt	centWave	matchedFilter	MZmine
0	25	75	65.5 ± 1.1	55.8 ± 1.1	61.5 ± 1
0	50	50	67.8 ± 0.7	58.3 ± 0.6	63.4 ± 0.8
0	75	25	67.1 ± 0.8	59.1 ± 0.8	64.2 ± 0.7

Tabelle 8.14: F-score für die Vereinigungsmenge von samen- und blattspezifischen Features aus dem Referenzdatensatz B. Der F-score wurde berechnet mittels der Recall-Werte in Tabelle 8.13 und der Precision-Werte in Tabelle 8.12.

Konzentrationsverhältnis (%)			Recall (%)		
Lösungsmittel	Samen	Blatt	centWave	matchedFilter	MZmine
0	25	75	47.7 ± 1.2	38.1 ± 1.2	41.9 ± 0.9
0	50	50	63.1 ± 1.7	50.6 ± 1.3	56.5 ± 1
0	75	25	76.4 ± 2.2	62 ± 2.3	70.7 ± 2.1
0	75	25	52.7 ± 1	44 ± 1.1	49.6 ± 0.9
0	50	50	68.7 ± 0.9	58.3 ± 1.2	65.3 ± 1.2
0	25	75	80.9 ± 0.8	70.5 ± 0.9	77.8 ± 0.6

Tabelle 8.15: Recall-Werte für samen- (oberer Teil) und blattspezifische (unterer Teil) Features aus dem Referenzdatensatz B.

Tabelle zur regelbasierten Annotation

Ion	x : Massendifferenz in [u]	n : Anzahl der Moleküle M im Ion	z : Ladungs-zustand
$[M+H]^+$	1.01	1	1
$[M+2H]^{2+}$	2.01	1	2
$[M+3H]^{3+}$	3.02	1	3
$[M+Na]^+$	22.99	1	1
$[M+2Na]^{2+}$	45.98	1	2
$[M+K]^+$	38.96	1	1
$[M+2K]^{2+}$	77.93	1	2
$[M+H+Na]^{2+}$	24.00	1	2
$[M+H+K]^{2+}$	39.97	1	2
$[M+Na+K]^{2+}$	61.95	1	2
$[M+NH_4]^+$	18.03	1	1
$[2M+H]^+$	1.01	2	1
$[2M+Na]^+$	22.99	2	1
$[2M+K]^+$	38.96	2	1
$[2M+H+Na]^{2+}$	24.00	2	2
$[2M+H+K]^{2+}$	39.97	2	2
$[3M+H]^+$	1.01	3	1
$[3M+Na]^+$	22.99	3	1
$[3M+K]^+$	38.96	3	1
$[3M+H+Na]^{2+}$	24.00	3	2
$[3M+H+K]^{2+}$	39.97	3	2
$[M+H\text{-}CH_4SO]^+$	-62.99	1	1
$[M+H\text{-}CH_4S]^+$	-47.00	1	1
$[M+H\text{-}C_6H_{10}O_5]^+$	-161.04	1	1
$[M+H\text{-}C_6H_{10}O_4]^+$	-145.05	1	1

Tabelle 8.16: Vorgegebene Massendifferenzen von Addukten und Fragmenten für die regelbasierte Annotation

Annotationsergebnisse auf dem MM14 Datensatz

Substanz Nr.	m/z	Retentionszeit	Intensität	Manuelle Annotation	Ergebnis des Algorithmus	Abgeleitete Molekülmasse	KG
1	135.05	280.43	97554	$[C8H7O2]^+$			1
1	153.06	280.43	4207	$[M+H]^+$	$[M+H]^+$	152.05 *	1
1	175.04	280.43	7468	$[M+Na]^+$	$[M+Na]^+$		1
1	197.02	280.76	1015	$[M+2Na-H]^+$			1
2	229.08	540.8	2428	$[C14H13O3]^+$			2
2	253.05	540.8	1968	$[M-CH_3O]^+$			2
2	270.05	540.8	5016	$[M-CH_2]^+$			2
2	285.07	540.8	644461	$[M+H]^+$	$[M+H]^+$		2
2	307.06	540.8	56425	$[M+Na]^+$	$[M+Na]^+$		2
2	591.12	541.14	8628	$[2M+Na]^+$	$[2M+Na]^+$	284.07	2
2	607.09	540.8	5465	$[2M+K]^+$	$[2M+K]^+$		2
2	891.16	540.8	583	$[3M+K]^+$	$[3M+K]^+$		2
3	117.04	278.07	995	$[C8H5O]^+$			3
3	145.03	278.07	20534	$[C9H5O2]^+$			3
3	149.06	278.07	3561	$[C9H9O2]^+$			3
3	177.05	278.07	31096	$[C10H9O3]^+$			3
3	195.06	278.07	1925	$[M+H]^+$	$[M+H]^+$		3
3	217.05	277.74	1704	$[M+Na]^+$	$[M+Na]^+$		3
3	233.01	278.07	3541	$[M+K]^+$	$[M+K]^+$	194.05	3
3	427.07	278.07	1897	$[2M+K]^+$	$[2M+K]^+$		3
3	621.13	278.07	435	$[3M+K]^+$	$[3M+K]^+$		3
4	118.08	383.03	1402	$[C5H12NO2]^+$			4
4	130.07	383.03	63197	$[C9H8N]^+$			4
4	158.06	383.37	1904	$[C10H8NO]^+$			4
4	258.13	383.03	1222	$[M+H-H2O]^+$			4
4	275.14	383.03	109028	$[M+H]^+$	$[M+H]^+$		4
4	297.12	383.03	84230	$[M+Na]^+$	$[M+Na]^+$		4
4	571.25	383.03	5746	$[2M+Na]^+$	$[2M+Na]^+$	274.13	4
4	587.22	383.03	10707	$[2M+K]^+$	$[2M+K]^+$		4
4	861.35	383.03	1126	$[3M+K]^+$	$[3M+K]^+$		4
4	593.23	383.03	3174	$[2M+2Na-H]^+$			4
4	316.16	383.71	2242	$[M+H+CH3CN]^+$			4

Fortsetzung auf nächster Seite

Substanz Nr.	*m/z*	Retentionszeit	Intensität	Manuelle Annotation	Ergebnis des Algorithmus	Abgeleitete Molekülmasse	KG
5	117.06	401.54	6026	$[C8H7N^{\circ}]^+$			5
5	118.06	402.21	773	$[C8H8N]^+$			-
5	130.07	401.54	45997	$[C9H8N]^+$			5
5	155.06	401.87	6635	$[C10H7N2]^+$			5
5	157.08	401.87	4016	$[M+H]^+$	$[M+H]^+$	156.07 *	5
5	179.06	401.54	992	$[M+Na]^+$	$[M+Na]^+$		5
6	117.06	300.61	5335	$[C8H7N^{\circ}]^+$			6
6	118.07	300.61	28975	$[C8H8N]^+$			6
6	144.05	300.61	899	[C9H6NO+]			6
6	146.06	300.61	8929	$[M+H]^+$	$[M+H]^+$	145.05 *	6
6	168.04	300.61	3893	$[M+Na]^+$	$[M+Na]^+$		6
7	287.05	425.42	236785	$[M+H]^+$	$[M+H]^+$	286.05	7
7	309.04	425.08	6257	$[M+Na]^+$	$[M+Na]^+$		7
7	611.05	425.42	1738	$[2M+K]^+$	$[2M+K]^+$		7
8	119.04	193.63	838	$[C5H3N4]^+$			-
8	136.06	193.3	7189	$[C5H6N5]^+$			-
8	148.06	192.96	26370	$[C6H6N5]^+$			8
8	173.07	192.96	2545	$[C8H7N5^{\circ}]^+$			8
8	188.09	192.96	6262	$[C9H10N5]^+$			8
8	198.08	192.96	526	$[C10H8N5]^+$			8
8	216.09	192.96	93990	$[M+H]^+$	$[M+H]^+$	215.08 *	8
8	238.07	192.62	8830	$[M+Na]^+$	$[M+Na]^+$		8
9	119.05	256.88	3623	$[C8H7O]^+$			9
9	147.05	256.88	18839	$[C9H7O2]^+$			9
9	165.05	256.54	910	$[M+H]^+$			-
9	203	256.88	1700	$[M+K]^+$			-
9	367.05	256.88	1913	$[2M+K]^+$			9
10	124.11	83.3	14362	$[C8H6D4N]^+$			10
10	125.11	83.64	65365	$[C8H5D5N]^+$			10
10	171.12	83.3	2700	$[C9H7D5NO2]^+$			10
11	135.05	43.93	10695	$[C8H7O2]^+$			11
11	152.07	43.93	670	$[M+H]^+$			11

Fortsetzung auf nächster Seite

Substanz Nr.	m/z	Retentionszeit	Intensität	Manuelle Annotation	Ergebnis des Algorithmus	Abgeleitete Molekülmasse	KG
12	127.04	415.33	419	$[C6H7O3]^+$			-
12	169.05	415.66	16494	$[C8H9O4]^+$			12
12	275.09	415.66	33013	$[M+H]^+$	$[M+H]^+$	274.08	12
12	297.07	415.66	19565	$[M+Na]^+$	$[M+Na]^+$		12
12	313.04	415.66	6567	$[M+K]^+$	$[M+K]^+$		12
12	431.1	415.66	634	$[3M+H+K]^{2+}$	$[3M+K+H]^{2+}$		12
12	571.15	415.66	2286	$[2M+Na]^+$	$[2M+Na]^+$		12
12	587.12	415.66	5371	$[2M+K]^+$	$[2M+K]^+$		12
13	107.05	327.52	1165	$[C7H7O]^+$			13
13	169.05	327.52	11247	$[C8H9O4]^+$			13
13	275.09	327.52	79327	$[C15H15O5]^+$			13
13	437.14	327.52	5550	$[M+H]^+$	$[M+H]^+$	436.13	13
13	448.12	327.86	892	$[2M+H+Na]^{2+}$	$[2M+H+Na]^{2+}$		13
13	456.11	327.52	21348	$[2M+H+K]^{2+}$	$[2M+K+H]^{2+}$		13
13	459.12	327.52	95600	$[M+Na]^+$	$[M+Na]^+$		13
13	475.09	327.52	11457	$[M+K]^+$	$[M+K]^+$		13
13	895.26	327.52	1444	$[2M+Na]^+$	$[2M+Na]^+$		13
13	911.23	327.19	1196	$[2M+K]^+$	$[2M+K]^+$		13
14	129.06	273.7	681	$[C6H9O3]^+$			-
14	147.07	273.36	428	$[C6H11O4]^+$			14
14	303.05	273.7	247427	$[C15H11O7]^+$			14
14	465.1	273.7	108543	$[C21H21O12]^+$			14
14	611.16	273.36	52776	$[M+H]^+$	$[M+H]^+$	610.15	14
14	630.12	273.7	29001	$[2M+H+K]^{2+}$	$[2M+K+H]^{2+}$		14
14	633.14	273.7	78195	$[M+Na]^+$	$[M+Na]^+$		14
14	649.11	273.7	5119	$[2M+2K]^{2+}$	$[2M+2K]^{2+}$		14
14	935.2	273.36	941	$[3M+H+K]^{2+}$	$[3M+K+H]^{2+}$		14

Tabelle 8.17: Ergebnisse der regel- und korrelationsbasierten Annotation auf dem MM14 Datensatz. Die Spalte KG bezeichnet die Zugehörigkeit des Features zu einer Korrelationsgruppe. Die Nummerierungen in dieser Spalte wurde nachträglich auf die in der ersten Spalte verwendete Nummerierung der Substanzen angepasst. In den mit * markierten Fällen wurden auch konkurrierende Annotationen mit Clusterionen vorgeschlagen (siehe Abschnitt 4.5). Dargestellt ist jeweils der erste Vorschlag, der die hier richtige Annotation als Quasi-Molekülionen enthält.

Substanz Nr.	Bezeichnung	Summenformel	Monoisotopische Molekülmasse
1	Anissäure	$C_8H_8O_3$	152.05
2	Biochanin A	$C_{16}H_{12}O_5$	284.07
3	Ferulsäure	$C_{10}H_{10}O_4$	194.06
4	IAA-Valin	$C_{15}H_{18}N_2O_3$	274.13
5	Indolacetonitril	$C_{10}H_8N_2$	156.07
6	Indolcarbaldehyd	C_9H_7NO	145.05
7	Kaempferol	$C_{15}H_{10}O_6$	286.05
8	Kinetin	$C_{10}H_9N_5O$	215.08
9	p-Coumarsäure	$C_9H_8O_3$	164.05
10	Phenylalanin-d5	$C_9H_6D_5NO_2$	170.11
11	Phenylglycin	$C_8H_9NO_2$	151.06
12	Phloretin	$C_{15}H_{14}O_5$	274.08
13	Phlorizin	$C_{21}H_{24}O_{10}$	436.14
14	Rutin	$C_{27}H_{30}O_{16}$	610.15

Tabelle 8.18: Summenformeln und monoisotopische Molekülmassen der 14 Substanzen im Datensatz MM14.

Substanz Nr.	Bezeichnung	Anzahl zusätzlicher Features in der Korrelations-Gruppe	Anzahl zusätzlich annotierter Features	Anzahl zusätzlich erkannter Molekülmassen
1	Anissäure	-	-	-
2	Biochanin A	30	2	1
3	Ferulsäure	10	2	1
4	IAA-Valin	69	10	5
5	Indolacetonitril	5	-	-
6	Indolcarbaldehyd	9	-	-
7	Kaempferol	18	-	-
8	Kinetin	2	-	-
9	p-Coumarsäure	3	-	-
10	Phenylalanin-d5	-	-	-
11	Phenylglycin	-	-	-
12	Phloretin	17	-	-
13	Phlorizin	43	8	4
14	Rutin	52	6	3

Tabelle 8.19: Anzahl zusätzlicher Features (d.h. nicht durch die manuelle Annotation erfasst) in den von der Korrelationsanalyse gefundenen Feature-Gruppen. In der Menge dieser zusätzlichen Features wurden durch die regelbasierte Annotation teilweise weitere Ionen annotiert, deren berechnete Molekülmasse jedoch nicht zur der als Marker verwendeten Substanz passt.

Retentionszeitabweichungen der Datensätze M1 und M2

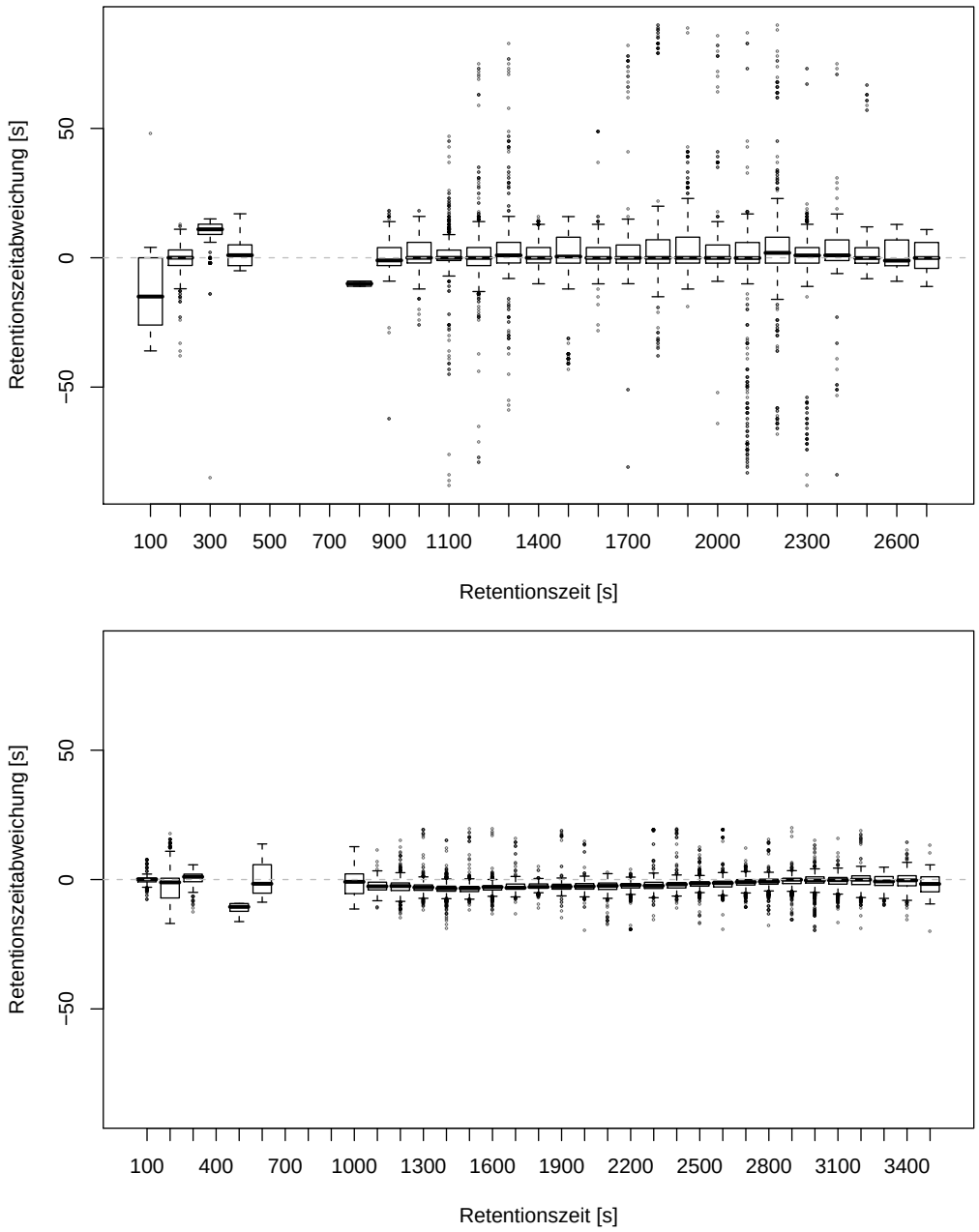

Abbildung 8.1: Box-Whiskers-Darstellung der Retentionszeitabweichungen in den Referenz-Datensätzen von M1 (oben) und M2 (unten). Für jedes Konsensus-Feature aus einem Referenz-Datensatz wurden die Differenzen der Retentionszeiten der einzelnen Features zur Retentionszeit des Features aus der ersten Messung darin gebildet. Diese Differenzen wurde zur Darstellung in Intervalle von 100 s gruppiert. Zur besseren Vergleichbarkeit wurden die y-Achsen mit derselben Skalierung versehen.

Alignment-Recall und -Precision Werte

Datensatz	MZmine	OpenMS	XAlign	XCMS	
				ohne Retentionszeitkorrektur	mit Retentionszeitkorrektur
M1					
$\text{Recall}_{\text{Align}}$	0.89	0.87	0.88	**0.95**	0.92
$\text{Precision}_{\text{Align}}$	0.74	0.69	0.70	0.66	0.73
M2					
$\text{Recall}_{\text{Align}}$	**0.98**	0.93	0.93	0.97	**0.98**
$\text{Precision}_{\text{Align}}$	**0.84**	0.79	0.79	0.58	0.78

Tabelle 8.20: $\text{Recall}_{\text{Align}}$ und $\text{Precision}_{\text{Align}}$-Werte für die Alignments der Datensätze M1 und M2.

Laufzeiten der Alignments

Datensatz	MZmine	OpenMS	XAlign	XCMS	
				ohne Retentionszeitkorrektur	mit Retentionszeitkorrektur
M1	20	4.4	51	0.9	1.4
M2	44	8.7	35	5.5	5.8
Total	64	13.1	86	6.4	7.2

Tabelle 8.21: Laufzeiten für die Alignments der Datensätze M1 und M2 in Minuten.

Literatur

[Aberg et al. 2008] Aberg, K.M., R. Torgrip, J. Kolmert, I. Schuppe-Koistinen und J. Lindberg (2008). *Feature detection and alignment of hyphenated chromatographic-mass spectrometric data Extraction of pure ion chromatograms using Kalman tracking*. J Chromatogr A, 1192:139–146.

[Aharoni et al. 2002] Aharoni, A., C. Ric de Vos, H. Verhoeven, C. Maliepaard, G. Kruppa, R. Bino und D. Goodenowe (2002). *Nontargeted metabolome analysis by use of Fourier Transform Ion Cyclotron Mass Spectrometry*. OMICS, 6:217–234.

[Amster 1996] Amster, I.J. (1996). *Fourier transform mass spectrometry*. J. Mass Spectrom., 31:1325–1337.

[Andreev et al. 2003] Andreev, V.P., T. Rejtar, H.-S. Chen, E. Moskovets, A. Ivanov und B. Karger (2003). *A Universal Denoising and Peak Picking Algorithm for LC-MS Based on Matched Filtration in the Chromatographic Time Domain*. Analytical Chemistry, 75(22):6314–6326.

[Ardrey 2003] Ardrey, Robert E. (2003). *Liquid Chromatography Mass Spectrometry: An Introduction (Analytical Techniques in the Sciences (AnTs) *)*. Wiley, 1. Aufl.

[Baldwin 2004] Baldwin, Michael A. (2004). *Protein Identification by Mass Spectrometry: Issues to be Considered*. Mol Cell Proteomics, 3(1):1–9.

[Balogh 2004] Balogh, Michael P. (2004). *Debating resolution and mass accuracy*. LC GC NORTH AMERICA, 17(3):152. `http://www.lcgceurope.com/lcgceurope/data/articlestandard//lcgceurope/112004/88264/article.pdf`.

[Baran et al. 2006] Baran, Richard, H. Kochi, N. Saito, M. Suematsu, T. Soga, T. Nishioka, M. Robert und M. Tomita (2006). *MathDAMP: a package for differential analysis of metabolite profiles*. BMC Bioinformatics, 7(1):530.

[Bayliss und Lashin 2006] Bayliss, Mark A. und V. Lashin (2006). *Why is Automating the Determination of Molecular Ions Using Automated Approaches so Hard and How Might it be Used?*. Poster at ASMS 2006, Seattle, WA, USA. `http://www.acdlabs.com/download/publ/2006/asms06_automating.pdf`.

[Beckmann et al. 2008] Beckmann, Manfred, D. Parker, D. P. Enot, E. Duval und J. Draper (2008). *High-throughput, nontargeted metabolite fingerprinting using nominal mass flow injection electrospray mass spectrometry*. Nat. Protocols, 3(3):486–504.

[Bellew et al. 2006] Bellew, Matthew, M. Coram, M. Fitzgibbon, M. Igra, T. Randolph, P. Wang, D. May, J. K. Eng, R. Fang, C. Lin, J. Chen, D. Goodlett, J. Whiteaker, A. G. Paulovich und M. McIntosh (2006). *A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS.*. Bioinformatics (Oxford, England), 22(15):1902–1909.

[Blatter 2003] Blatter, Christian (2003). *Wavelets. Eine Einführung.*. Vieweg, 2. Aufl.

[Boccard et al. 2007] Boccard, Julien, E. Grata, A. Thiocone, J.-Y. Gauvrit, P. Lanteri, P.-A. Carrupt, J.-L. Wolfender und S. Rudaz (2007). *Multivariate data analysis of rapid LC-TOF/MS experiments from Arabidopsis thaliana stressed by wounding*. Chemometrics and Intelligent Laboratory Systems, 86:189–197.

[Böcker und Lipták 2005] Böcker, Sebastian und Z. Lipták (2005). *Efficient mass decomposition*. In: *SAC '05: Proceedings of the 2005 ACM symposium on Applied computing*, S. 151–157, New York, NY, USA. ACM.

[van den Bogaert et al. 1994] Bogaert, Bas van den, H. F. Boelens und H. C. Smit (1994). *Quantification of overlapping chromatographic peaks using a matched filter*. Chemometrics and Intelligent Laboratory Systems, 25(2):297–311.

[Bruker 2007] Bruker (2007). *Target Analysis User Manual, Version 1.1*. User Manual, Bruker Daltonik GmbH.

[Budzikiewicz und Schäfer 2005] Budzikiewicz, Herbert und M. Schäfer (2005). *Massenspektrometrie. Eine Einführung*. Wiley-VCH, 5. Aufl.

[Bure und Lange 2003] Bure, C. und C. Lange (2003). *Comparison of Dissociation of Ions in an Electrospray Source, or a Collision Cell in Tandem Mass Spectrometry*. Current Organic Chemistry, 7:1613–1624(12).

[Bylund et al. 2002] Bylund, D., R. Danielsson, G. Malmquist und K. E. Markides (2002). *Chromatographic alignment by warping and dynamic programming as a pre-processing tool for PARAFAC modelling of liquid chromatography mass spectrometry data*. J Chromatogr A, 961(2):237–244.

[Böcker et al. 2006] Böcker, Sebastian, M. C. Letzel, Z. Lipták und A. Pervukhin (2006). *Decomposing Metabolomic Isotope Patterns.*. In: Bucher, Philipp und B. M. E. Moret, Hrsg.: *WABI*, Bd. 4175 d. Reihe *Lecture Notes in Computer Science*, S. 12–23. Springer.

[Böttcher et al. 2007] Böttcher, C., E. Roepenack-Lahaye, E. Willscher, D. Scheel und S. Clemens (2007). *Evaluation of Matrix Effects in Metabolite Profiling Based on Capillary Liquid Chromatography Electrospray Ionization Quadrupole Time-of-Flight Mass Spectrometry*. Analytical Chemistry, 79(4):1507–1513.

[Böttcher et al. 2008] Böttcher, Christoph, E. von Röpenack-Lahaye, J. Schmidt, C. Schmotz, S. Neumann, D. Scheel und S. Clemens (2008). *Metabolome Analysis of Biosynthetic Mutants Reveals a Diversity of Metabolic Changes and Allows Identification of a Large Number of New Compounds in Arabidopsis*. Plant Physiol., 147(4):2107–2120.

[Chen et al. 2006] Chen, E.I., J. Hewel, B. Felding-Habermann und J. Yates (2006). *Large scale protein profiling by combination of protein fractionation and multidimensional protein identification technology (MudPIT)*. Mol. Cell Proteomics, 5:53–56.

[Chernushevich et al. 2001] Chernushevich, I.V., A. Loboda und B. Thomson (2001). *An introduction to quadrupole-time-of-flight mass spectrometry*. J Mass Spectrom, 36:849–865.

[Cole 1997] Cole, Richard B., Hrsg. (1997). *Electrospray Ionization Mass Spectrometry*. Wiley-Interscience, 1. Aufl.

[Conrad et al. 2006] Conrad, Tim O. F., A. Leichtle, A. Hagehülsmann, E. Diederichs, S. Baumann, J. Thiery und C. Schütte (2006). *Beating the Noise: New Statistical Methods for Detecting Signals in MALDI-TOF Spectra Below Noise Level.*. In: Berthold, Michael R., R. C. Glen und I. Fischer, Hrsg.: *CompLife*, Bd. 4216 d. Reihe *Lecture Notes in Computer Science*, S. 119–128. Springer.

[Cooks und Rockwood 1991] Cooks, R. G. und A. Rockwood (1991). *The 'Thomson'. A suggested unit for mass spectroscopists*. Rapid Communications in Mass Spectrometry, 5(2):93.

[Danielsson et al. 2002] Danielsson, R., D. Bylund und K. Markides (2002). *Matched filtering with background suppression for improved quality of base peak chromatograms and mass spectra in liquid chromatography - mass spectrometry*. Analytica Chimica Acta, (454):167–184.

[Daszykowski und Walczak 2006] Daszykowski, Michal und B. Walczak (2006). *Use and abuse of chemometrics in chromatography*. TrAC Trends in Analytical Chemistry, 25(11):1081–1096.

[Daubechies 1992] Daubechies, Ingrid (1992). *Ten lectures on wavelets*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

[De Vos et al. 2007] De Vos, R.C., S. Moco, A. Lommen, J. Keurentjes, R. Bino und R. Hall (2007). *Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry*. Nature Protocols, 2:778–791.

[Di Marco und Bombi 2001] Di Marco, V.B. und G. Bombi (2001). *Mathematical functions for the representation of chromatographic peaks*. J Chromatogr A, 931:1–30.

[Du et al. 2006] Du, Pan, W. A. Kibbe und S. M. Lin (2006). *Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching.*. Bioinformatics, 22(17):2059–2065.

[Dunn 2008] Dunn, Warwick B (2008). *Current trends and future requirements for the mass spectrometric investigation of microbial, mammalian and plant metabolomes*. Physical Biology, 5(1):011001 (24pp).

[Felinger 1998] Felinger, Attila (1998). *Data Analysis and Signal Processing in Chromatography (Data Handling in Science and Technology)*. Elsevier Science, 1 Aufl.

[Fiehn 2001] Fiehn, O. (2001). *Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks*. Comp. Funct. Genomics, 2:155–168.

[Fiehn et al. 2000] Fiehn, O., J. Kopka, P. Dörmann, T. Altmann, R. Trethewey und L. Willmitzer (2000). *Metabolite profiling for plant functional genomics*. Nature Biotechnology, 18:115.

[Gentleman et al. 2004] Gentleman, Robert C, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. H. Yang und J. Zhang (2004). *Bioconductor: Open software development for computational biology and bioinformatics*. Genome biology, 5:R80.

[Gibon et al. 2006] Gibon, Y., B. Usadel, O. Blaesing, B. Kamlage, M. Hoehne, R. Trethewey und M. Stitt (2006). *Integration of metabolite with transcript and enzyme activity profiling during diurnal cycles in Arabidopsis rosettes*. Genome Biol., 7:R76.

[Grabowski 2004] Grabowski, Barbara (2004). *Lexikon der Statistik: Mit ausführlichem Anwendungsteil*. Spektrum Akademischer Verlag, 1 Aufl.

[Grange und Sovocool 2008] Grange, A.H. und G. Sovocool (2008). *Automated determination of precursor ion, product ion, and neutral loss compositions and deconvolution of composite mass spectra using ion correlation based on exact masses and relative isotopic abundances*. Rapid Commun. Mass Spectrom., 22:2375–2390.

[Grange et al. 2006] Grange, A.H., M. Zumwalt und G. Sovocool (2006). *Determination of ion and neutral loss compositions and deconvolution of product ion mass spectra using an orthogonal acceleration time-of-flight mass spectrometer and an ion correlation program*. Rapid Commun. Mass Spectrom., 20:89–102.

[Gröpl et al. 2005] Gröpl, C., E. Lange, K. Reinert, O. Kohlbacher, M. Sturm, C. G. Huber, B. Mayr und C. Klein (2005). *Algorithms for the automated absolute quantification of diagnostic markers in complex proteomics samples*. In: Berthold, Michael, Hrsg.: *Procceedings of CompLife 2005*, Lecture Notes in Bioinformatics, S. 151–163. Springer, Heidelberg.

[Gross 2004] Gross, Jürgen H. (2004). *Mass Spectrometry: A Textbook*. Springer, Berlin, 2. Aufl.

[Guilhaus 1995] Guilhaus, M. (1995). *Principles and instrumentation in time of flight mass spectrometry*. J. Mass Spectrom., 30:1519 –1532.

[Görlach und Richmond 1999] Görlach, E. und R. Richmond (1999). *Discovery of Quasi-Molecular Ions in Electrospray Spectra by Automated Searching for Simultaneous Adduct Mass Differences*. Analytical Chemistry, 71(24):5557–5562.

[Hartuv und Shamir 2000] Hartuv, Erez und R. Shamir (2000). *A clustering algorithm based on graph connectivity*. In: *Information Processing Letters*, S. 175–181.

[Hoffmann und Stroobant 2001] Hoffmann, Edmond De und V. Stroobant (2001). *Mass Spectrometry: Principles and Applications*. John Wiley and Sons Ltd, 2. Aufl.

[Horai et al. 2008] Horai, Hisayuki, M. Arita und T. Nishioka (2008). *Comparison of ESI-MS Spectra in MassBank Database*. bmei, 2:853–857.

[Iijima et al. 2008] Iijima, Y., Y. Nakamura, Y. Ogata, K. Tanaka, N. Sakurai, K. Suda, T. Suzuki, H. Suzuki, K. Okazaki, M. Kitayama, S. Kanaya, K. Aoki und D. Shibata (2008). *Metabolite annotations based on the integration of mass spectral information*. Plant J., 54:949–962.

[Jaitly et al. 2006] Jaitly, N., M. Monroe, V. Petyuk, T. Clauss, J. Adkins und R. Smith (2006). *Robust Algorithm for Alignment of Liquid Chromatography-Mass Spectrometry Analyses in an Accurate Mass and Time Tag Data Analysis Pipeline*. Anal. Chem., 78(21):7397–7409.

[Kalman 1960] Kalman, R. E. (1960). *A New Approach to Linear Filtering and Prediction Problems*. Transactions of the ASME – Journal of Basic Engineering, (82 (Series D)):35–45.

[Kamber und Han 2006] Kamber, Micheline und J. Han (2006). *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2 Aufl.

[Katajamaa und Oresic 2007] Katajamaa, M. und M. Oresic (2007). *Data processing for mass spectrometry-based metabolomics*. J Chromatogr A, 1158:318–328.

[Katajamaa et al. 2005] Katajamaa, Mikko, J. Miettinen und M. Oresic (2005). *Processing methods for differential analysis of LC/MS profile data*. BMC bioinformatics, 6:179.

[Katajamaa und Oresic 2005] Katajamaa, Mikko und M. Oresic (2005). *Processing methods for differential analysis of LC/MS profile data*. BMC Bioinformatics, 6(1):179.

[Keller et al. 2008] Keller, B., J. Sui, A. Young und R. Whittal (2008). *Interferences and contaminants encountered in modern mass spectrometry*. Analytica Chimica Acta, 627(1):71–81.

[Kiencke et al. 2008] Kiencke, Uwe, M. Schwarz und T. Weickert (2008). *Signalverarbeitung: Zeit-Frequenz-Analyse und Schätzverfahren*. Oldenbourg.

[Kind und Fiehn 2007] Kind, Tobias und O. Fiehn (2007). *Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry*. BMC Bioinformatics, 8(1):105.

[Kohlbacher et al. 2007] Kohlbacher, Oliver, K. Reinert, C. Gröpl, E. Lange, N. Pfeifer, O. Schulz-Trieglaff und M. Sturm (2007). *TOPP–the OpenMS proteomics pipeline*. Bioinformatics, 23(2):191–197.

[Kolotilin et al. 2007] Kolotilin, I., H. Koltai, Y. Tadmor, C. Bar-Or, M. Reuveni, A. Meir, S. Nahon, H. Shlomo, L. Chen und I. Levin (2007). *Transcriptional profiling of high pigment-2dg tomato mutant links early fruit plastid biogenesis with its overproduction of phytonutrients*. Plant Physiol., 145:389–401.

[Kromidas 2006] Kromidas, Stavros, Hrsg. (2006). *HPLC richtig optimiert: Ein Handbuch für Praktiker*. Wiley-VCH, 1. Aufl.

[Lange et al. 2006] Lange, E., C. Gröpl, K. Reinert, O. Kohlbacher und A. Hildebrandt (2006). *High-accuracy peak picking of proteomics data using wavelet techniques.*. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, S. 243–254.

[Lange et al. 2007] Lange, Eva, C. Gröpl, O. Schulz-Trieglaff, A. Leinenbach, C. Huber und K. Reinert (2007). *A Geometric Approach for the Alignment of Liquid Chromatography-Mass Spectrometry Data*. In: *Proceedings of the 15th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) & 6th European Conference on Computational Biology (ECCB)*, S. i273–i281.

[Lange et al. 2008] Lange, Eva, R. Tautenhahn, S. Neumann und C. Gröpl (2008). *Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements*. BMC Bioinformatics, 9:375.

[Li et al. 2005] Li, Xiao-Jun, E. C. Yi, C. J. Kemp, H. Zhang und R. Aebersold (2005). *A Software Suite for the Generation and Comparison of Peptide Arrays from Sets of Data Collected by Liquid Chromatography-Mass Spectrometry*. Molecular & cellular proteomics : MCP, 4(9):1328–1340.

[Listgarten und Emili 2005] Listgarten, J. und A. Emili (2005). *Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry*. Molecular & cellular proteomics : MCP, 4:419–434.

[Listgarten et al. 2007] Listgarten, Jennifer, R. M. Neal, S. T. Roweis, P. Wong und A. Emili (2007). *Difference detection in LC-MS data for protein biomarker discovery*. Bioinformatics (Oxford, England), 23(2):e198–204.

[Lottspeich 2006] Lottspeich, Friedrich (2006). *Bioanalytik*. Spektrum, Akad. Verl., 2. Aufl.

[Mallet et al. 2006] Mallet, Claude R, E. Chambers, D. M. Diehl und J. R. Mazeo (2006). *A Study of Contributions from HPLC Vials To Ion Suppression/Enhancement in Electrospray Ionization*. `http://www.waters.com/webassets/cms/library/docs/wa43207.pdf`. Poster, Waters Corporation.

[MarkerView] MarkerView. *MarkerView^{TM} Version 1.1*. `https://products.appliedbiosystems.com/ab/en/US/adirect/ab?cmd=catNavigate2&catID=601522`. Applied Biosystems/MDS SCIEX.

[McLerran et al. 2008] McLerran, Dale F., Z. Feng, O. J. Semmes, L. Cazares und T. W. Randolph (2008). *Signal Detection in High-Resolution Mass Spectrometry Data*. Journal of Proteome Research, 7(1):276–285.

[Meffert und Hochmuth 2004] Meffert, Beate und O. Hochmuth (2004). *Werkzeuge der Signalverarbeitung. Grundlagen, Anwendungsbeispiele, Übungsaufgaben*. Pearson Studium.

[MetAlign 2006] MetAlign (2006). *MetAlign^{TM} Version 09.11.2006*. `http://www.pri.wur.nl/UK/products/MetAlign/`. RIKILT - Institute of Food Safety and Plant Research International.

[Meyer 2004] Meyer, Veronika R. (2004). *Praxis der Hochleistungs-Flüssigkeitschromatographie*. Wiley-VCH, 9. Aufl.

[Mihaleva et al. 2008] Mihaleva, V.V., O. Vorst, C. Maliepaard, H. Verhoeven, d. Vos, R.C.H., R. Hall und v. Ham, R.C.H.J. (2008). *Accurate mass error correction in liquid chromatography time-of-flight mass spectrometry based metabolomics*. `http://library.wur.nl/file/wurpubs/wurpublikatie_i00367362_001.pdf`.

[Miller und Denton 1986] Miller, P. E. und M. B. Denton (1986). *The quadrupole mass filter: basic operating concepts*. J. Chem. Ed., 63:617–622.

[Mueller et al. 2003] Mueller, Lukas A., P. Zhang und S. Y. Rhee (2003). *AraCyc: A Biochemical Pathway Database for Arabidopsis*. Plant Physiol., 132(2):453–460.

[Nordström et al. 2006] Nordström, A., G. O'Maille, C. Qin und G. Siuzdak (2006). *Nonlinear Data Alignment for UPLC-MS and HPLC-MS Based Metabolomics: Quantitative Analysis of Endogenous and Exogenous Metabolites in Human Serum*. Anal. Chem., 78(10):3289–3295.

[Prakash et al. 2006] Prakash, Amol, P. Mallick, J. Whiteaker, H. Zhang, A. Paulovich, M. Flory, H. Lee, R. Aebersold und B. Schwikowski (2006). *Signal Maps for Mass Spectrometry-based Comparative Proteomics*. Molecular & cellular proteomics : MCP, 5(3):423–432.

[Prince und Marcotte 2006] Prince, J.T. und E. Marcotte (2006). *Chromatographic Alignment of ESI-LC-MS Proteomics Data Sets by Ordered Bijective Interpolated Warping*. Anal. Chem., 78(17):6140–6152.

[Radulovic et al. 2004] Radulovic, D., S. Jelveh, S. Ryu, T. Hamilton, E. Foss, Y. Mao und A. Emili (2004). *Informatics Platform for Global Proteomic Profiling and Biomarker Discovery Using Liquid Chromatography-Tandem Mass Spectrometry*. Molecular & cellular proteomics : MCP, 3(10):984–997.

[Rijsbergen 1979] Rijsbergen, C. J. Van (1979). *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA.

[van Rijswick 1974] Rijswick, M.H.J. van (1974). *Adaptive program for high precision off-line processing of chromatograms*. Chromatographia, 7(9):491–501.

[Roepenack-Lahaye et al. 2004] Roepenack-Lahaye, E. von, T. Degenkolb, M. Zerjeski, M. Franz, U. Roth, L. Wessjohann, J. Schmidt, D. Scheel und S. Clemens (2004). *Profiling of Arabidopsis Secondary Metabolites by Capillary Liquid Chromatography Coupled to Electrospray Ionization Quadrupole Time-of-Flight Mass Spectrometry*. Plant Physiology, 134:548–559.

[Roessner et al. 2001] Roessner, U., A. Luedemann, D. Brust, O. Fiehn, T. Linke, L. Willmitzer und A. Fernie (2001). *Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems*. Plant Cell, 13:11–29.

[Sato et al. 2004] Sato, S., T. Soga, T. Nishioka und M. Tomita (2004). *Simultaneous determination of the main metabolites in rice leaves using capillary electrophoresis mass spectrometry and capillary electrophoresis diode array detection*. Plant J., 40:151–163.

[Senko et al. 1995] Senko, M.W., S. Beu und F. McLafferty (1995). *Determination of Monoisotopic Masses and Ion Populations for Large Biomolecules from Resolved Isotopic Distributions*. Journal of the American Society for Mass Spectrometry, 6(4):229–233(5).

[Smith et al. 2006] Smith, C. A., E. J. Want, G. O'Maille, R. Abagyan und G. Siuzdak (2006). *XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification*. Anal. Chem., 78(3):779–787.

[Stolt et al. 2006] Stolt, R., R. Torgrip, J. Lindberg, L. Csenki, J. Kolmert, I. Schuppe-Koistinen und S. Jacobsson (2006). *Second-Order Peak Detection for Multicomponent High-Resolution LC/MS Data*. Analytical Chemistry, 78(4):975–983.

[Sturm et al. 2008] Sturm, Marc, A. Bertsch, C. Gröpl, A. Hildebrandt, R. Hussong, E. Lange, N. Pfeifer, O. Schulz-Trieglaff, A. Zerck, K. Reinert und O. Kohlbacher (2008). *OpenMS - An open-source framework for mass spectrometry*. BMC bioinformatics, 9:163. `http://www.openms.de`.

[Swartz 2005] Swartz, M. E. (2005). *Ultra performance liquid chromatography (UPLC): An introduction*. Separation Science Redefined (Supplement to LC-GC), 8(14).

[Tautenhahn et al. 2007] Tautenhahn, Ralf, C. Böttcher und S. Neumann (2007). *Annotation of LC/ESI-MS Mass Signals*. In: Hochreiter, Sepp und R. Wagner, Hrsg.: *BIRD*, Bd. 4414 d. Reihe *Lecture Notes in Computer Science*, S. 371–380. Springer.

[Tautenhahn et al. 2008] Tautenhahn, Ralf, C. Böttcher und S. Neumann (2008). *Highly Sensitive Feature Detection For High Resolution LC/MS*. BMC Bioinformatics, 9(508).

[Tikunov et al. 2005] Tikunov, Y., A. Lommen, C. d. Vos, H. Verhoeven, R. Bino, R. Hall und A. Bovy (2005). *A novel approach for nontargeted data analysis for metabolomics. Large-scale profiling of tomato fruit volatiles.*. Plant Physiol, 139(3):1125–37.

[Tolstikov und Fiehn 2002] Tolstikov, V.V. und O. Fiehn (2002). *Analysis of highly polar compounds of plant origin: combination of hydrophilic interaction chromatography and electrospray ion trap mass spectrometry*. Anal. Biochem., 301:298–307.

[Tolstikov et al. 2003] Tolstikov, V.V., A. Lommen, K. Nakanishi, N. Tanaka und O. Fiehn (2003). *Monolithic silica-based capillary reversed-phase liquid chromatography/electrospray mass spectrometry for plant metabolomics*. Anal. Chem., 75:6737–6740.

[Trieglaff et al. 2008] Trieglaff, Ole S., N. Pfeifer, C. Gropl, O. Kohlbacher und K. Reinert (2008). *LC-MSsim - a simulation software for Liquid Chromatography Mass Spectrometry data*. BMC Bioinformatics, 9(1).

[Tuszynski 2008] Tuszynski, Jarek (2008). *The caTools Package*. `http://cran.r-project.org/web/packages/caTools/caTools.pdf`. R-Paket caTools, Version 1.9.

[Wang et al. 2007] Wang, Pei, H. Tang, M. P. Fitzgibbon, M. Mcintosh, M. Coram, H. Zhang, E. Yi und R. Aebersold (2007). *A statistical method for chromatographic alignment of LC-MS data*. Biostatistics (Oxford, England), 8(2):357–367.

[Wang et al. 2008] Wang, S.-C., C.-J. Lin, S.-M. Chiang und S.-N. Yu (2008). *Tailoring Noise Frequency Spectrum between Two Consecutive Second Derivative Filtering Procedures to Improve Liquid Chromatography-Mass Spectrometry Determinations*. Analytical Chemistry, 80(6):2097–2104.

[Wang et al. 2006] Wang, Shau-Chun, S.-M. Chiang und C.-M. Huang (2006). *Parametric studies of matched filters to enhance the signal-to-noise ratios of LC-MS-MS peaks*. Analytica Chimica Acta, 556(1):201–207.

[Ward et al. 2003] Ward, J.L., C. Harris, J. Lewis und M. Beale (2003). *Assessment of 1H NMR spectroscopy and multivariate analysis as a technique for metabolite fingerprinting of Arabidopsis thaliana*. Phytochemistry, 62:949–957.

[Werner et al. 2008] Werner, E., J. F. Heilier, C. Ducruix, E. Ezan, C. Junot und J. C. Tabet (2008). *Mass spectrometry for the identification of the discriminating signals from metabolomics: Current status and future trends.*. Journal of chromatography. B, Analytical technologies in the biomedical and life sciences, 871(2):143–163.

[Wolters et al. 2001] Wolters, D. A., M. P. Washburn und J. R. Yates (2001). *An automated multidimensional protein identification technology for shotgun proteomics.*. Anal. Chem., 73(23):5683–5690.

[Yu 2002] Yu, Hao (2002). *Rmpi: Parallel Statistical Computing in R*. R News, 2(2):10–14. `http://CRAN.R-project.org/doc/Rnews/Rnews_2002-2.pdf`.

[Zhang et al. 2005] Zhang, X., J. Asara, J. Adamec, M. Ouzzani und A. K. Elmagarmid (2005). *Data pre-processing in liquid chromatography / mass spectrometry-based proteomics*. Bioinformatics (Oxford, England), 21(21):4054–4059.

[Zhang und McElvain 1999] Zhang, Z. und J. McElvain (1999). *Optimizing Spectroscopic Signal-to-Noise Ratio in Analysis of Data Collected by a Chromatographic/Spectroscopic System*. Analytical Chemistry, 71(1):39–45.

Erklärung

Hiermit erkläre ich, dass ich diese Arbeit selbständig und ohne fremde Hilfe verfasst habe. Ich habe keine anderen als die von mir angegebenen Quellen und Hilfsmittel benutzt. Die den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen sind als solche kenntlich gemacht worden. Ich habe mich bisher nicht um den Doktorgrad beworben.

Halle (Saale), den 11. Dezember 2008 Ralf Tautenhahn

Lebenslauf

Persönliche Daten

Name	Ralf Tautenhahn
geboren am	04. März 1976
in	Schlema
Staatsangehörigkeit	deutsch
Familienstand	ledig

Schulausbildung

9/1982 - 8/1991	117. Polytechnische Oberschule Dresden
9/1991 - 6/1994	Gymnasium Dresden-Plauen, Abschluss Abitur

Universitätsausbildung

10/1997 - 09/2000	Studium der Biologie an der Technischen Universität Dresden Abschluss mit Vordiplom
10/2000 - 12/2004	Studium der Bioinformatik an der Martin-Luther-Universität Halle Wittenberg Abschluss als Diplom-Bioinformatiker
seit 03/2005	wissenschaftlicher Mitarbeiter am Leibniz-Institut für Pflanzenbiochemie, Halle (Saale) Arbeitsgruppe Massenspektrometrie & Bioinformatik

Halle (Saale), den 11. Dezember 2008 Ralf Tautenhahn